Statistical Methods for Variable Selection in Causal Inference

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Brandon Lee D. Koch

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Julian Wolfson, David Vock

July, 2018

Acknowledgements

I am extremely grateful to my advisors, Julian Wolfson and David Vock, as their guidance, assistance, and unwavering support have helped me tremendously throughout my time in graduate school. I am very grateful to Jim Hodges for his advice and assistance that considerably strengthened my dissertation.

I want to thank Lan Liu for taking the time to review and help improve my dissertation. I also want to thank Laura Boehm Vock for greatly improving Chapter 3 of this dissertation.

Dedication

This dissertation is dedicated to my loving friends and family, but most especially, to my wife, daughter, and parents.

Abstract

Estimating the causal effect of a binary intervention or action (referred to as a "treatment") on a continuous outcome is often an investigator's primary goal. Randomized trials are ideal for estimating causal effects because randomization eliminates selection bias in treatment assignment. However, randomized trials are not always ethically or practically possible, and observational data must be used to estimate the causal effect of treatment. Unbiased estimation of causal effects with observational data requires adjustment for confounding variables that are related to both the outcome and treatment assignment. Adjusting for all measured covariates in a study protects against bias, but including covariates unrelated to outcome may increase the variability of the estimated causal effect. Standard variable selection techniques aim to maximize predictive ability of a model for the outcome and are used to decrease variability of the estimated causal effect, but they ignore covariate associations with treatment and may not adjust for important confounders weakly associated to outcome. We propose two approaches for estimating causal effects that simultaneously consider models for both outcome and treatment assignment. The first approach is a variable selection technique for identifying confounders and predictors of outcome using an adaptive group lasso approach that simultaneously performs coefficient selection, regularization, and estimation across the treatment and outcome models. In the second approach, two methods are proposed that simultaneously model outcome and treatment assignment using a Bayesian formulation with spike and slab priors on each covariate coefficient; the Spike and Slab Causal Estimator (SSCE) aims to achieve minimum bias of the causal effect estimator while Bilevel SSCE (BSSCE) aims to minimize its mean squared error. We also propose TEHTrees, a new method that combines matching and conditional inference trees to characterize treatment effect heterogeneity. One of its main virtues is that, by employing formal hypothesis testing procedures in constructing the tree, TEHTrees preserves the Type I error rate.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Estimating the causal effect of a binary intervention or action (referred to as a "treatment") on a continuous outcome is often an investigator's primary goal. Randomized trials are ideal for estimating causal effects because randomization eliminates selection bias in treatment assignment. However, randomized trials are not always ethically or practically possible, and observational data must be used to estimate the causal effect of treatment. When using observational data to estimate the casual effect of treatment, many methods require either modeling the mean outcome conditional on the predictors and the treatment (e.g., regression modeling), or specifying a treatment allocation model (e.g., inverse probability weighting (IPW) and propensity score matching), or both (e.g., doubly robust methods) (Lunceford and Davidian, 2004). Methods that rely on only one such model require that the model be specified correctly and adjust for at least all confounders – variables associated with both treatment and outcome – for consistent estimation of the causal treatment effect. Doubly robust methods, however, fit both an outcome and a treatment model and require only one of them be specified correctly with all confounders for consistent treatment effect estimation.

One approach is to include all available covariates in the specified model(s) to avoid biased estimation. However, including many variables unrelated to outcome and treatment could inflate the variance of the effect estimator. Hence, when there are a large number of possible confounders, some type of variable selection is desirable to achieve unbiased, efficient estimation. VanderWeele and Shpitser (2011) propose a confounder selection criterion that controls for any covariate that is either a cause of treatment or outcome.

Though efficiency may improve by including covariates related only to outcome, as shown by Brookhart et al. (2006) for IPW estimators and de Luna, Waernbaum, and Richardson (2011) for non-parametric estimators of the average causal effect (ACE), including all causes of treatment or outcome can still be sub-optimal as these studies also suggest efficiency may decrease when controlling for variables that are related to the treatment but not the outcome.

Variable selection methods (e.g., backward variable selection, lasso) based only on the outcome (treatment, respectively) model are popular in practice, but because these methods ignore the relationship between treatment (outcome) and covariates, these methods tend to under-select confounding variables weakly related to the outcome (treatment) but strongly associated with the treatment (outcome). Vansteelandt, Bekaert, and Claeskens (2012) argue that omitting such variables in estimators of the ACE not only introduces bias but also underestimates the uncertainty of the ACE and propose a method based on a focused information criterion which aims at minimizing the mean squared error of the treatment effect estimator.

There has been work to adapt traditional variable selection techniques, which focus on covariates with the greatest predictive ability of treatment or outcome, to jointly select covariates related to treatment and outcome. van der Laan and Gruber (2010) propose a doubly robust semi-parametric method that solves an efficient influence curve equation that is a function of the outcome and treatment models by utilizing numerous data adaptive machine learning algorithms to select variables in a stepwise fashion for the propensity score. Ertefaie, Asgharian, and Stephens (2015) proposed a two-step variable selection method which selects variables using a penalized likelihood in the first step and then separately estimates the causal treatment effect in the second step using a doubly robust regression estimator. A limitation of this method, however, is that it may not select an important confounder if its association with the outcome and treatment have opposite signs; this can occur when the value of the coefficient in the outcome and treatment likelihoods are similar in magnitude.

Bayesian model averaging (BMA) proposes taking a weighted average of the effect estimates across models with different covariates included where the weights are determined by the posterior model probability (Raftery, Madigan, & Hoeting, 1997). However, like traditional variable selection, standard BMA tends to prioritize models which include covariates

strongly associated with outcome and may assign significant weight to models that include only a subset of the necessary confounders, resulting in biased treatment effect estimation (Crainiceanu, Dominici, & Parmigiani, 2008). Crainiceanu et al. (2008) introduce a two-stage BMA method that forces strong predictors of treatment that are identified in a first stage to be included in the outcome model in a second stage, and then strong predictors of outcome that are not identified in stage one are identified in stage two. Wang, Parmigiani, & Dominici (2012) propose Bayesian Adjustment for Confounding (BAC), a Bayesian model averaging method on the outcome model with an informative prior obtained separately from the treatment model (see Zigler & Dominici (2014) for related Bayesian methods that select variables for propensity score estimation). BAC contains a prior dependence parameter, ω, ranging from 1 to ∞ that links the treatment model to the outcome model. If $\omega = \infty$, all covariates with associations in the treatment model are forced into the outcome model, whereas $\omega = 1$ treats the two models independently and is equivalent to standard BMA on the outcome model.

Two approaches have been predominantly used to select ω: (1) setting ω equal to ∞ as the default and (2) selecting ω data-adaptively to minimize mean squared error (MSE) or other criterion. Each approach is problematic. Setting $\omega = \infty$ targets the set of covariates associated with treatment or outcome. However, we observe that BAC with $\omega = \infty$ can have high inclusion probabilities for irrelevant covariates that are unrelated to outcome and treatment, particularly in smaller sample sizes. This can lead to inefficient estimators in moderate sample sizes compared to other variable selection approaches which target the same set of covariates (i.e., all variables associated with treatment or outcome). Furthermore, for a given sample size, selecting all covariates related to treatment and outcome may not be the set which leads to the most efficient estimator of the average causal effect.

However, developing a data-adaptive approach to select ω to minimize MSE has proved challenging. Lefebvre, Atherton, & Talbot (2014) proposed using cross validation or the bootstrap to choose ω with the aim of minimizing MSE of the treatment effect estimator, but they found that the performance of these procedures was sensitive to the underlying data generating mechanism and suggested that alternative approaches should be investigated. Even if cross validation or the bootstrap could be reliably used to choose ω, such methods can be computationally intensive with large datasets. Further, BAC requires calculating the Bayesian Information Criterion at each posterior draw, which cannot be calculated when

the number of potential confounders in the model exceeds the sample size.

In Chapter 2, we propose GLiDeR (Group Lasso and Doubly Robust Estimation), a treatment effect estimator which uses a modified adaptive group lasso approach (Yuan and Lin, 2006) to perform simultaneous coefficient regularization and estimation for the treatment and outcome models. Our method is more efficient than standard (doubly robust) backward selection methods and is competitive with the two-stage BMA estimator proposed by Cefalu et al. (2017). However, unlike the two-stage BMA estimator, our proposed method is computationally feasible with a very large number of covariates including cases where the number of covariates is larger than the sample size.

Chapter 3, we propose the Spike and Slab Causal Estimator (SSCE) and Bilevel SSCE (BSSCE), novel Bayesian methods that simultaneously consider models for outcome and treatment and use spike and slab priors on the covariate coefficients to encourage variable selection based on associations in both the outcome and treatment models. SSCE aims to minimize treatment effect bias by controlling only for covariates that are related to outcome or treatment (and removing irrelevant ones), while BSSCE adjusts for the subset of the covariates which minimize MSE of the treatment effect estimator. The proposed methods, which are adapted from the formulation of the Bayesian group lasso with spike and slab priors (Xu & Ghosh, 2015), are implemented using fast Gibbs samplers that perform well with a large number of covariates, even when the number of covariates is greater than the sample size.

Chapters 2 and 3 concern variable selection for the estimation of the average causal effect. However, many policy and medical decisions are informed by heterogeneous treatment effects that are found in the results of randomized studies. For example, BiDil is the first drug to be FDA-approved for a single racial group because study results suggest it is beneficial for African-Americans with congestive heart failure. However, if BiDil is not truly beneficial for African-Americans with congestive heart failure (i.e., if a Type I error was made), then production and marketing of the drug would prove to be very wasteful and the potential side effects of BiDil would likely cause harm with no benefit for patients who have received the drug.

Traditional approaches to characterizing treatment effect heterogeneity have centered around regression modeling with interaction terms between the treatment/intervention indicator and covariates. If treatment is randomized, the magnitude and statistical significance

of the interaction term can be used to assess the degree of treatment effect heterogeneity. However, when there are multiple interactions, regression models can quickly become difficult to interpret. Also, conclusions about the presence of interaction depend on the way in which interaction terms are specified, so that important subgroups may not be identified. Lastly, considering multiple interactions can inflate Type I error, but it is common not to adjust the significance level for multiple testing, which leads to anti-conservative inference (Green and Kern, 2012).

Recently, there has been increasing interest in developing techniques to estimate heterogeneous treatment effects using more flexible models that can "automatically" detect subgroups of interest. Athey and Imbens (2016) developed Causal Tree, which uses a single regression tree to recursively partition the data into homogeneous subgroups that have similar treatment effects and a similar subset of covariate values. Wager and Athey (2017) extend Causal Tree to Causal Forests, which averages treatment effect estimates over many Causal Trees. Green and Kern (2012) propose another method that uses multiple regression trees, called Bayesian Adaptive Regression Trees (BART), which automatically detects nonlinear relationships and interactions to describe treatment effect heterogeneity. While many of these approaches provide flexible subgroup identification, they typically do not address the issue of inflation of Type I error.

In Chapter 4, we propose a novel method, called Treatment Effect Heterogeneity Trees (TEHTrees), for characterizing treatment effect heterogeneity. TEHTrees combines matching with conditional inference trees (Hothorn and others, 2012). One of its main virtues is that, by employing formal testing procedures in constructing the tree, TEHTrees preserves Type I error. In simulation studies comparing TEHTrees and Causal Tree, the Type I error rate is less than the desired 0.05 level in all considered scenarios using TEHTrees, but is greater than 0.15 in all scenarios using Causal Tree (and far greater in some cases). Though the power (defined as the probability of splitting on the variable with true heterogeneous treatment effects) is slightly larger using Causal Tree compared to TEHTrees in our simulations with continuous covariates, the power with binary covariates is actually greater using TEHTrees compared to that of Causal Tree.

Chapter 2

Covariate selection with group lasso and doubly robust estimation of causal effects

2.1 Introduction

Estimating the causal effect of a binary intervention or action (referred to as a "treatment") on a continuous outcome is often an investigator's primary goal. Randomized trials are ideal for estimating causal effects because randomization eliminates selection bias in treatment assignment. However, randomized trials are not always ethically or practically possible, and observational data must be used to estimate the causal effect of treatment. When using observational data to estimate the casual effect of treatment, many methods require either modeling the mean outcome conditional on the predictors and the treatment (e.g., regression modeling), or specifying a treatment allocation model (e.g., inverse probability weighting (IPW) and propensity score matching), or both (e.g., doubly robust methods) (Lunceford and Davidian, 2004). Methods that rely on only one such model require that the model be specified correctly and adjust for at least all confounders – variables associated with both treatment and outcome – for consistent estimation of the causal treatment effect. Doubly robust methods, however, fit both an outcome and a treatment model and require only one of them be specified correctly with all confounders for consistent treatment effect estimation.

One approach is to include all available covariates in the specified model(s) to avoid biased estimation. However, including many variables unrelated to outcome and treatment could inflate the variance of the effect estimator. Hence, when there are a large number of possible confounders, some type of variable selection is desirable to achieve unbiased, efficient estimation. VanderWeele and Shpitser (2011) propose a confounder selection criterion that controls for any covariate that is either a cause of treatment or outcome. Though efficiency may improve by including covariates related only to outcome, as shown by Brookhart et al. (2006) for IPW estimators and de Luna, Waernbaum, and Richardson (2011) for non-parametric estimators of the average causal effect (ACE), including all causes of treatment or outcome can still be sub-optimal as these studies also suggest efficiency may decrease when controlling for variables that are related to the treatment but not the outcome. Variable selection methods (e.g., backward variable selection, lasso) based only on the outcome (treatment, respectively) model are popular in practice, but because these methods ignore the relationship between treatment (outcome) and covariates, these methods tend to under-select confounding variables weakly related to the outcome (treatment) but strongly associated with the treatment (outcome). Vansteelandt, Bekaert, and Claeskens (2012) argue that omitting such variables in estimators of the ACE not only introduces bias but also underestimates the uncertainty of the ACE and propose a method based on a focused information criterion which aims at minimizing the mean squared error of the treatment effect estimator.

There has been work to adapt traditional variable selection techniques, which focus on covariates with the greatest predictive ability of treatment or outcome, to jointly select covariates related to treatment and outcome. van der Laan and Gruber (2010) propose a doubly robust semi-parametric method that solves an efficient influence curve equation that is a function of the outcome and treatment models by utilizing numerous data adaptive machine learning algorithms to select variables in a stepwise fashion for the propensity score. Ertefaie et al. (2015) proposed a two-step variable selection method which selects variables using a penalized likelihood in the first step and then separately estimates the causal treatment effect in the second step using a doubly robust regression estimator. A limitation of this method, however, is that it may not select an important confounder if its association with the outcome and treatment have opposite signs; this can occur when the value of the

coefficient in the outcome and treatment likelihoods are similar in magnitude. Wang, Parmigiani, and Dominici (2012) propose Bayesian adjustment for confounding (BAC), a method linking the models for treatment and outcome with a dependence parameter. Cefalu et al. (2017) take a similar approach to BAC by developing a two-stage Bayesian model averaged (BMA) doubly robust method that introduces a prior dependence between a covariates' inclusion in the propensity score and the outcome model that is designed to identify the set of potential confounders based on their association with both treatment and outcome by forcing variables included in the propensity score to be a subset of those included in the outcome model. Despite improved efficiency over standard methods, these approaches must estimate a posterior distribution on some model class, which is typically done using measures (e.g., BIC) that cannot handle situations when the number of covariates is larger than the sample size. Even when the number of predictors is less than the sample size, these methods can be computationally intensive when the number of predictors is large as they must explore all possible treatment and outcome model spaces; with even a modest number of covariates, say 20, over 2 million (2^{21}) models must be considered. Moreover, since treatment effect estimates are weighted linear combinations across many models, there is no feature selection and interpretation of covariate effects is difficult.

In this chapter, we propose GLiDeR (Group Lasso and Doubly Robust Estimation), a treatment effect estimator which uses a modified adaptive group lasso approach (Yuan and Lin, 2006) to perform simultaneous coefficient regularization and estimation for the treatment and outcome models. Our method is more efficient than standard (doubly robust) backward selection methods and is competitive with the two-stage BMA estimator proposed by Cefalu et al. (2017). However, unlike the two-stage BMA estimator, our proposed method is computationally feasible with a very large number of covariates including cases where the number of covariates is larger than the sample size.

We set up the problem and introduce the group lasso in Chapter 2.2. In Chapter 2.3, we formulate GLiDeR and summarize the estimation technique. Chapter 2.4 provides theoretical justification and asymptotic results for GLiDeR. In Chapter 2.5 we present simulation scenarios demonstrating the finite-sample behavior of GLiDeR, and Chapter 2.6 provides an application to an observational registry of lung transplant recipients. We conclude in Chapter 2.7.

2.2 Preliminaries

2.2.1 Doubly robust estimation of treatment effects

The causal effect of binary treatment A on continuous outcome Y is of interest. Letting $Y(a)$ denote the possibly counterfactual outcome for a randomly selected person if assigned treatment $A = a$, the ACE is $\Delta := E[Y(1) - Y(0)]$. When A is randomized, the vector of potential outcomes $\{Y(0), Y(1)\}$ is independent of A. Given data (Y_i, A_i) on independent subjects $i = 1, \ldots, n$, $\hat{\Delta}_{ran} = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1-A_i) Y_i}{\sum_{i=1}^n 1-A_i}$ is a consistent estimator of Δ.

In an observational study, $\{Y(0), Y(1)\}$ may depend on A and $\hat{\Delta}_{ran}$ may then be inconsistent for Δ. However, it may be reasonable to assume that treatment assignment is *ignorable* and has positive probability (*positivity*) given observed covariates, i.e., that there exist covariates $\mathbf{X} = \{X_1, \ldots, X_m\}$ such that $A \perp Y(a)|\mathbf{X}$ and $P(A = a|\mathbf{X}) > 0$, for $a = \{0, 1\}$, in which case Δ is consistent. We can then postulate a regression model $\mu(A, \mathbf{X}; \alpha)$ for $E(Y|A, \mathbf{X})$. If $\mu(A, \mathbf{X}; \alpha_0) = E(Y|A, \mathbf{X})$ for some α_0 (i.e., the outcome model is correctly specified), then given a consistent estimator $\hat{\alpha}$ of α_0, the estimator $\hat{\Delta}_{reg} = \frac{1}{n} \sum_{i=1}^n [\mu(1, \mathbf{X}_i; \hat{\alpha}) - \mu(0, \mathbf{X}_i; \hat{\alpha})]$ is consistent for Δ (Lunceford and Davidian, 2004). If $\mu(A, \mathbf{X}; \alpha) \neq E(Y|A, \mathbf{X})$, then $\hat{\Delta}_{reg}$ may be inconsistent for Δ.

Let $\pi(\mathbf{X}; \gamma)$ be a postulated regression model for the conditional probability of treatment, $P(A = 1|\mathbf{X})$. If $\pi(\mathbf{X}; \gamma_0) = P(A = 1|\mathbf{X})$ for some γ_0 (i.e., the treatment model is correctly specified), then a consistent estimator of Δ is the inverse probability weighted (IPW) estimator, $\hat{\Delta}_{IPW} = \frac{1}{n} \sum_{i=1}^n \left[\frac{A_i Y_i}{\pi(\mathbf{X}_i; \hat{\gamma})} - \frac{(1-A_i) Y_i}{1-\pi(\mathbf{X}_i; \hat{\gamma})} \right]$, where $\hat{\gamma}$ is any consistent estimator of γ_0 (Lunceford and Davidian, 2004). If $\pi(\mathbf{X}; \gamma) \neq P(A = 1|\mathbf{X})$, then $\hat{\Delta}_{IPW}$ may be an inconsistent estimator for Δ.

To address the problem of model misspecification, various authors have proposed doubly robust estimators, which require specification of both an outcome and propensity score model but require only one of them to be correctly specified to yield a consistent estimator for Δ (Lunceford and Davidian, 2004). One such doubly robust estimator is

$$\hat{\Delta}_{DR} = \frac{1}{n} \sum_{i=1}^n \left[\frac{A_i Y_i}{\pi(\mathbf{X}_i; \hat{\gamma})} - \frac{(1-A_i) Y_i}{1-\pi(\mathbf{X}_i; \hat{\gamma})} - \left[\frac{A_i - \pi(\mathbf{X}_i; \hat{\gamma})}{\pi(\mathbf{X}_i; \hat{\gamma})} \right] \mu(1, \mathbf{X}_i; \hat{\alpha}) - \left[\frac{A_i - \pi(\mathbf{X}_i; \hat{\gamma})}{1-\pi(\mathbf{X}_i; \hat{\gamma})} \right] \mu(0, \mathbf{X}_i; \hat{\alpha}) \right]. \quad (2.1)$$

The preceding has assumed that the observed covariates $\mathbf{X}$ are exactly those required to achieve ignorability, i.e., $\mathbf{X}$ is precisely the set of confounders of the treatment-outcome

relationship. However, in practice, we may have access to a large set of covariates $\mathbf{V} \supset \mathbf{X}$ which are candidates for inclusion in the outcome and treatment models. While the estimators mentioned remain consistent if covariates from $\mathbf{V} \setminus \mathbf{X}$ are added to the propensity and outcome models (in addition to $\mathbf{X}$), in Chapter 2.4.1, we show that including covariates related only to the outcome can decrease the variance – while adding covariates associated with only the treatment can increase the variance – of the doubly robust estimator. In Chapter 2.3, we introduce GLiDeR, a procedure for performing simultaneous variable selection in treatment and outcome models that targets confounders and predictors of only outcome.

2.2.2 The Group Lasso

Our approach to simultaneous variable selection in the outcome and treatment models uses a modified version of the group lasso (Yuan and Lin, 2006) – a regularization method that acts like the lasso (Tibshirani, 1996) on grouped covariates by forcing all coefficients of each group of variables to be either all zero or all nonzero. We briefly introduce the group lasso technique for a general regression model before describing our particular modification of it in the next subchapter.

Let $\mathbf{M}$ be the $q \times 1$ vector of covariates corresponding to a regression model for some response variable R, and let ξ be the vector of associated regression coefficients. In the group lasso, we assume that $\mathbf{M}$ is partitioned into K groups $\{\mathbf{M}_1, \ldots, \mathbf{M}_K\}$; the corresponding blocks of ξ are denoted by $\xi^{(1)}, \ldots, \xi^{(K)}$. For a general loss function $\Phi(R, \mathbf{M}; \xi)$, the group lasso estimator of ξ is $\hat{\xi}_{GL}(\lambda) = \underset{\xi}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Phi(R_i, \mathbf{M}_i; \xi) + \lambda \sum_{k=1}^{K} w_k \|\xi^{(k)}\|_2$, where $\lambda > 0$ is the penalty parameter and $w_k \geq 0$ is the penalty weight for group k. A common choice for each group weight w_k is $\sqrt{c_k}$, where c_k is the cardinality of group k (i.e., the number of elements in $\mathbf{M}_k$). If we do not want to penalize a specific group, for example an intercept, we let the corresponding w_k equal 0. There is no closed form solution to $\hat{\xi}_{GL}(\lambda)$, but several algorithms exist, including the groupwise majorization descent (GMD) algorithm proposed by Yang and Zou (2015), for finding solutions to this convex optimization problem.

2.3 GLiDeR

2.3.1 Notation

Let $\mathbf{V}_i = \{V_{1i}, \ldots, V_{pi}\}$ denote subject i's vector of measured covariates, with p possibly large in relation to the sample size n; for the remainder of the chapter, we suppress i in our notation except where necessary. We assume as in Chapter 2.2.1 *ignorability* and *positivity* of treatment assignment given $\mathbf{V}$. Let outcome and propensity models for $E(Y|A, \mathbf{V})$ and $P(A = 1|\mathbf{V})$ be defined by $f[\mu(A, \mathbf{V}; \alpha)] = \alpha_1 V_1 + \cdots + \alpha_p V_p + \alpha_{p+1} + \alpha_{p+2} A$, and $g[\pi(\mathbf{V}; \gamma)] = \gamma_1 V_1 + \ldots \gamma_p V_p + \gamma_{p+1}$, and let $\Phi_{out}(Y, A, \mathbf{V}; \alpha)$ and $\Phi_{trt}(A, \mathbf{V}; \gamma)$ denote the outcome and treatment loss functions used to fit these models. In many doubly robust treatment effect estimation problems, f is taken to be the identity function and g is the logit function, so that the outcome and treatment models represent linear and logistic regression. In this case, Φ_{out} is the squared error loss and Φ_{trt} is proportional to the binomial negative log-likelihood.

Anticipating the group lasso approach in the next subchapter, we will let $\beta = (\alpha, \gamma)$ and define $p + 3$ groups of this vector: $\beta_1 = (\alpha_1, \gamma_1), \ldots, \beta_p = (\alpha_p, \gamma_p)$, $\beta_{p+1} = \alpha_{p+1}$, $\beta_{p+2} = \gamma_{p+1}$, and $\beta_{p+3} = \alpha_{p+2}$. Note that, for $k = 1, \ldots, p$, β_k is a group of coefficients corresponding to the covariate V_k in the outcome and treatment model, respectively. Our setup differs from the typical one for group lasso, as our groupings correspond to the same covariate appearing in two different regression models, as opposed to sets of related but distinct covariates within the same regression model. Covariate transformations may be included by adding the necessary elements to $\mathbf{V}$ and grouping the coefficients of the transformed covariates with those of the untransformed versions. In this manuscript, we do not consider interactions between covariates; while including interactions poses no technical challenges, it is not clear to which group the corresponding columns in the design matrix for the interaction should belong.

2.3.2 Simultaneous variable selection for the treatment and outcome models

To perform simultaneous variable selection between the treatment and outcome models, we propose to solve a group lasso-like problem with the following characteristics:

(1) The loss function is taken to be the sum of the loss functions for the treatment and

outcome models,

$$\Phi_{sum}(Y, A, \mathbf{V}; \beta) = \Phi_{out}(Y, A, \mathbf{V}; \alpha) + \Phi_{trt}(A, \mathbf{V}; \gamma). \tag{2.2}$$

(2) We use the penalty term

$$P(\beta) = \lambda \sum_{k=1}^{K} W_k \|\beta_k\|_2 \equiv \lambda \sum_{k=1}^{p} W_k \sqrt{\alpha_k^2 + \gamma_k^2}$$

so each summand corresponds to the coefficients associated with a single covariate; $\lambda > 0$ is the penalty parameter and W_k is a weight term with $W_k = 0$ for $k > p$, that is we do not penalize the intercepts in the treatment and outcome models and the main effect of treatment in the outcome model. We discuss the choice of W_k in Chapter 2.3.3. Unlike the usual group lasso setup, where related covariates in the same model are jointly penalized, GLiDeR groups together the coefficients corresponding to the same covariate across the treatment and outcome models. This strategy forces covariates to enter and leave the models simultaneously.

Our simultaneous variable selection procedure therefore consists of solving

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Phi_{sum}(Y_{std_i}, A_i, \mathbf{V}_i; \beta) + \lambda \sum_{k=1}^{p} W_k \|\beta_k\|_2, \tag{2.3}$$

where $Y_{std} = Y/sd(Y)$ is used instead of Y in Equation (2.3) so the scale of $\Phi_{out}(Y, \mathbf{V}; \alpha)$ (and $\Phi_{sum}(Y, A, \mathbf{V}; \beta)$, by definition) does not depend on the measurement unit of continuous Y; and, therefore, estimation of β is not affected by the scale of Y. We also assume the covariates used in (2.3) are standardized so that the penalty is invariant to scale. Given a solution $\hat{\beta}(\lambda)$ of (2.3), we can plug $\hat{\beta}(\lambda) = \{\hat{\alpha}(\lambda), \hat{\gamma}(\lambda)\}$ into $\mu\{A, \mathbf{V}; \hat{\alpha}(\lambda)\}$ and $\pi\{\mathbf{V}; \hat{\gamma}(\lambda)\}$ to obtain an estimate of Δ at each λ, denoted $\hat{\Delta}_{DR}(\lambda)$, using Equation (2.1). As with the usual (group) lasso, the degree of variable selection is controlled by λ. We discuss choosing λ in Chapter 2.3.4.

2.3.3 Choosing W_k

Because our goal is to minimize the mean squared error of the treatment effect estimator and, therefore, encourage selection of covariates which are associated with the outcome

(which includes confounders) and discourage selection of covariates which are related only to the treatment or unrelated to both outcome and treatment, we propose to set $W_k = \frac{\sqrt{2}}{|v_k|}$, where the numerator corresponds to the cardinality of group k (the default group penalty weight under the general group lasso formulation) and $v_k \neq 0$ is an estimate of the regression coefficient in the outcome model for covariate k from a "full" model. When $p + 2 \leq n$, one can set v_k to be the ordinary least squares estimate for covariate k as obtained when fitting the full outcome model. In cases where $p + 2 > n$, one choice is the least squares coefficient estimate of covariate k with the ridge penalty. With transformations in the outcome or treatment model, W_k can be defined by setting the numerator equal to the square root of the total number of predictors in the outcome and treatment models that correspond to covariate k, and v_k equal to the l_2 norm of the corresponding estimated coefficients. When the weights vary based on the strength of the association between the covariate and outcome, we refer to GLiDeR as "adaptive." Ertefaie et al. (2015) also propose an adaptive weight so that the magnitude of the penalty on each coefficient is proportional to its contribution to the outcome model but is different than the adaptive weight proposed here as it depends on the least squares (or ridge) estimates of the coefficients in both the outcome and treatment models. Like the general group lasso, if we do not want to penalize a specific group, which is often the case for intercepts or the main effect of treatment in the outcome model, we set the corresponding group weights W_k to zero. This proposed group weight strongly penalizes covariates that are not associated with the outcome (i.e., when $|v_k|$ is small) even if they are strongly associated with the treatment. Hence, the adaptive approach used in GLiDeR is aimed to select covariates associated with only the outcome or confounders that are related to both treatment and outcome (i.e., covariates related to the treatment should be selected only if they are also associated with the outcome).

2.3.4 Choosing λ

Equation (2.3) defines the solution $\hat{\beta}(\lambda)$ as a function of λ. We propose to choose λ by applying cross-validation to the outcome model, since doing so further encourages the (desirable) selection of predictors associated with the outcome. Generalized cross-validation (GCV) or *k-fold* cross-validation (kCV) is typically used to select the tuning parameter λ in lasso-like problems. Since we consider both outcome and treatment model loss functions in Equation (2.3) but wish to apply GCV to only the outcome model, the usual GCV statistic

requires a modification (kCV is straightforward). For the general group lasso, the GCV statistic at a particular value λ is $\frac{RSS}{(1-df/n)^2}$, where RSS is the residual sum of squares and df $= \sum_{j=1}^{K} I\left(||\hat{\xi}_{aGL_j}|| > 0\right) + \sum_{j=1}^{K} \frac{||\hat{\xi}_{aGL_j}||}{||\tilde{\xi}_{aGL_j}||}(d_j - 1)$, where $\hat{\xi}_{aGL_j}$ and $\tilde{\xi}_{aGL_j}$ are the adaptive group lasso and least squares estimators of the jth group of coefficients, respectively, for groups $j = 1, \ldots, K$ with group sizes d_j. For GLiDeR, since we want a model selection procedure for the outcome model only, we take the residual sum of squares from the outcome model to use in the numerator and use only the parts of $\hat{\beta}(\lambda)$ corresponding to coefficients from the outcome model (denoted $\hat{\alpha}(\lambda)$) in the denominator, yielding the following modified GCV statistic (noting we have p "groups" of size 2 and 2 terms – the intercept α_{p+1} and treatment main effect α_{p+2} – in the outcome model that are unpenalized in separate groups of size 1):

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha}_1(\lambda)V_{1i} - \cdots - \hat{\alpha}_p(\lambda)V_{pi} - \hat{\alpha}_{p+1}(\lambda) - \hat{\alpha}_{p+2}(\lambda)A_i)^2}{\left(1 - \left(2 + \sum_{j=1}^{K} I\left(||\hat{\alpha}_j(\lambda)|| > 0\right) + \sum_{j=1}^{K} \frac{|\hat{\alpha}_j(\lambda)|}{|v_j|}\right)/n\right)^2}. \tag{2.4}$$

Then $\lambda^* = \min_{\lambda} GCV(\lambda)$ is the "optimal" λ. GCV is computationally advantageous since it only needs to be computed from the data once (as opposed to kCV, which needs to be computed an additional k times), and also demonstrates slightly better performance than kCV in the simulation scenarios considered in Chapter 2.5 (see Table 2.7 for results comparing GCV and kCV). We thus recommend using the GCV statistic in Equation (2.4) to select λ. Our final estimate of Δ is then $\hat{\Delta}_{DR}(\lambda^*)$ in Equation (2.1).

2.3.5 Implementation

To summarize, we now list the steps involved in implementing the GLiDeR procedure.

Step 1 - Define covariate groups: Group outcome and treatment model predictors (assumed to be standardized) as described in Chapter 2.3.1. For each group k, compute group weights W_k as described in Chapter 2.3.3. For groups k that represent intercepts in either model or the treatment main-effect term in the outcome model, let the corresponding group weight W_k be zero. Scale Y by its marginal standard deviation, as discussed in Chapter 2.3.2.

Step 2 - Apply the modified group lasso. Define a sequence of λ values $\lambda_1, \ldots, \lambda_L$, such that $\lambda_1 > \lambda_2 > \cdots > \lambda_L \geq 0$ with initial value λ_1 defined to be the smallest value λ such that all predictors have zero coefficients, except the terms with group weights (W_k)

equal to zero. For $l = 1, \ldots, L$, apply the GMD algorithm described in Yang and Zou (2015). Chapter 2.8.1 gives details of adapting this algorithm for this application.

Step 3 - Select the final model and estimate the doubly robust treatment effect. Use Equation (2.4) to compute $GCV(\lambda_l)$ for $l = 1, \ldots, L$ and let $\lambda^* = \underset{\lambda_l}{min}\, GCV(\lambda_l)$. Plug $\hat{\beta}(\lambda^*) = (\hat{\alpha}, \hat{\gamma})$ into $\mu(A, \mathbf{V}; \hat{\alpha})$ and $\pi(\mathbf{V}; \hat{\gamma})$ and obtain an estimate of Δ using Equation (2.1).

2.4 Asymptotic results

2.4.1 Efficient variable sets for doubly robust estimators

We begin by showing that including covariates related only to the treatment may increase – while including those related only to the outcome may decrease – the asymptotic variance of the doubly robust estimator, thereby justifying the covariate sets GLiDeR seeks to identify.

We consider doubly robust estimators in the class of (2.1) and focus attention on estimating $\mu_1 = E\{Y(1)\}$; ideas are similar for estimating $E\{Y(0)\}$ and, therefore, the ACE, $\Delta = E\{Y(1) - Y(0)\}$. A doubly robust estimator for μ_1 in the class of (2.1) is

$$\hat{\Delta}_{DR,\mu_1} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i Y_i}{\pi(\mathbf{V}_i;\hat{\gamma})} - \frac{A_i - \pi(\mathbf{V}_i;\hat{\gamma})}{\pi(\mathbf{V}_i;\hat{\gamma})}\mu(1, \mathbf{V}_i; \hat{\alpha})\right\}. \tag{2.5}$$

Let α^* and γ^* be the values of α and γ so that $E\left\{\frac{d}{d\alpha}\Phi_{out}(Y, A, \mathbf{V}; \alpha)|_{\alpha=\alpha^*}\right\} = 0$ and $E\left\{\frac{d}{d\gamma}\Phi_{trt}(A, \mathbf{V}; \gamma)|_{\gamma=\gamma^*}\right\} = 0$. If the models are correctly specified, then α^* and γ^* are the "true" values of the parameters, and if incorrectly specified, then these are the "least false" parameters. To investigate the effect of including certain types of covariates in $\hat{\Delta}_{DR,\mu_1}$, we will consider the case that $\gamma = \gamma^*$ and $\alpha = \alpha^*$ are known, so that $\pi(\mathbf{V}; \gamma^*)$ and $\mu(A, \mathbf{V}; \alpha^*)$ are known functions of $\mathbf{V}$. When one or both models are correctly specified, we can then show the asymptotic variance of $\sqrt{n}\hat{\Delta}_{DR,\mu_1}$ is $\Sigma_{DR}(\mathbf{V}) = Var\{Y(1)\} + E\left[\frac{1-\pi(\mathbf{V};\gamma^*)}{\pi(\mathbf{V};\gamma^*)}E\left\{(Y(1) - \mu(1, \mathbf{V}; \alpha^*))^2|\mathbf{V}\right\}\right]$, which follows from the iterated conditional variance formula. Now consider a new covariate, Z_1, and assume that $\gamma^*_{Z_1} \neq 0$ and $\alpha^*_{Z_1} = 0$, where $\gamma^*_{Z_1}$ and $\alpha^*_{Z_1}$ are the true/least false regression coefficients for Z_1 in the treatment and outcome models, respectively. That is, Z_1 is conditionally (given $\mathbf{V}$) related to treatment, but conditionally unrelated to the outcome.

Then the asymptotic variance of $\sqrt{n}\hat{\Delta}_{DR,\mu_1}$ with $\mathbf{V}$ and Z_1 is

$$\Sigma_{DR}(\mathbf{V}, Z_1) = Var\{Y(1)\} + E\left[\frac{1-\pi(\mathbf{V}, Z_1; \gamma^*, \gamma^*_{z_1})}{\pi(\mathbf{V}, Z_1; \gamma^*, \gamma^*_{z_1})} E\left[(Y(1) - \mu(1, \mathbf{V}, Z_1; \alpha^*, \alpha^*_{z_1}))^2 | \mathbf{V}, Z_1\right]\right].$$

Assuming the propensity score follows a logistic regression model, we can find $\frac{1-\pi(\mathbf{V}, Z_1; \gamma^*, \gamma^*_{z_1})}{\pi(\mathbf{V}, Z_1; \gamma^*, \gamma^*_{z_1})} = \frac{1-\pi(\mathbf{V}; \gamma^*)}{\pi(\mathbf{V}; \gamma^*)} \exp(-\gamma^*_{z_1} Z_1)$. Then since $\alpha^*_{Z_1} = 0$,

$$\Sigma_{DR}(\mathbf{V}, Z_1) = Var\{Y(1)\} + E\left[\exp(-\gamma^*_{z_1} Z_1) \frac{1-\pi(\mathbf{V}; \gamma^*)}{\pi(\mathbf{V}; \gamma^*)} E\left[(Y(1) - \mu(1, \mathbf{V}; \alpha^*))^2 | \mathbf{V}, Z_1\right]\right].$$

If Z_1 is independent of $\mathbf{V}$ and $Y(1)$, then $\Sigma_{DR}(\mathbf{V}, Z_1) = E\left[\exp(-\gamma^*_{z_1} Z_1)\right] \Sigma_{DR}(\mathbf{V})$. If Z_1 is centered at zero and normally distributed, we know that $E[\exp(-\gamma^*_{z_1} Z_1)] = \exp\left(\gamma^{*2}_{z_1}\sigma^2/2\right)$,where σ^2 is the variance of Z_1. Then as Z_1 is associated with treatment, $\gamma^*_{z_1} \neq 0$ and we have $\exp\left(\gamma^{*2}_{z_1}\sigma^2/2\right) > 1$ so that $\Sigma_{DR}(\mathbf{V}, Z_1) > \Sigma_{DR}(\mathbf{V})$, and $\Sigma_{DR}(\mathbf{V}, Z_1)$ gets larger as $|\gamma^*_{z_1}|$ increases. The same derivation holds for covariate $Z_1^* = Z_1 - E(Z_1|\mathbf{V})$ (i.e., Z_1^* is independent of V and $\alpha^*_{Z_1^*} = 0$) when Z_1 and $\mathbf{V}$ are multivariate normal and dependent.

If we instead consider an irrelevant covariate, Z_2, which follows the same assumptions as Z_1 except that it is conditionally unrelated to the propensity score so that $\gamma^*_{z_2} = \alpha^*_{z_2} = 0$, then a similar argument can be made to show that $\Sigma_{DR}(\mathbf{V}, Z_2) = \Sigma_{DR}(\mathbf{V})$. Lastly, consider covariate Z_3, which is assumed to be conditionally related to outcome but conditionally unrelated to treatment. Then the asymptotic variance of $\sqrt{n}\hat{\Delta}_{DR,\mu_1}$ with $\mathbf{V}$ and Z_3 is

$$\Sigma_{DR}(\mathbf{V}, Z_3) = Var\{Y(1)\} + E\left(\frac{1-\pi(\mathbf{V}; \gamma^*)}{\pi(\mathbf{V}; \gamma^*)} E\left[\{Y(1) - \mu(1, \mathbf{V}, Z_3; \alpha^*, \alpha^*_{z_3})\}^2 | \mathbf{V}, Z_3\right]\right),$$

which follows assuming $\gamma^*_{z_3} = 0$ (the truth). When the outcome model is correctly specified,

$$E\left[\{Y(1) - \mu(1, \mathbf{V}, Z_3; \alpha^*, \alpha^*_{z_3})\}^2 | \mathbf{V}, Z_3\right] < E\left[\{Y(1) - \mu(1, \mathbf{V}; \alpha^*)\}^2 | \mathbf{V}\right],$$

which implies $\Sigma_{DR}(\mathbf{V}, Z_3) < \Sigma_{DR}(\mathbf{V})$ when the regression error does not depend on covariates (i.e., homoscedastic); when the outcome model is misspecified, $\Sigma_{DR}(\mathbf{V}, Z_3) < \Sigma_{DR}(\mathbf{V})$ under homoscedasticity if prediction of $Y(1)$ via $\mu(1, \mathbf{V}, Z_3; \alpha^*, \alpha^*_{z_3})$ is improved over $\mu(1, \mathbf{V}; \alpha^*)$.

In practice, α and γ are not known and must be estimated. M-estimation techniques can be used to derive the asymptotic variance of the doubly robust estimator when α and

γ are estimated, but such derivations do not provide any obvious expressions that reveal the effect on the asymptotic variance of the doubly robust estimator after adding Z_1, Z_2, or Z_3 when one of the models is misspecified; when both models are correctly specified, the asymptotic variance is the same regardless of whether α and γ are known or estimated.

2.4.2 Targeted covariate sets and double robustness of GLiDeR

We now show GLiDeR can asymptotically recover the set of confounders and covariates related only to outcome, while excluding irrelevant variables and covariates related only to treatment. Specifically, GLiDeR selects a covariate V_j provided $\alpha^*_{V_j} \neq 0$. The key theorem is described here; the proof, which uses concentration inequalities from Blazère, Loubes, and Gamboa (2014), appears in Chapter 2.8.2.

Assume no transformations or interactions between covariates (i.e., all groups are of size 2) so that $W_k = \frac{\sqrt{2}}{|v_k|}$ as in Chapter 2.3.3. Assume further that there are no confounders such that $\alpha^*_{V_j} = 0$ if the outcome model is misspecified (i.e., the covariate has no linear association with the outcome, which rules out symmetric quadratic or periodic relationships). Then the group weight (W_k) tends to infinity for any covariate that is unrelated to the outcome, which allows covariates that are irrelevant or related only to treatment to be asymptotically excluded from GLiDeR. We can then prove the following theorem:

Theorem 1. *Assume the number of covariates p and sample size n are such that $\frac{\log(2p)}{n} \leq 1$. Also assume the Group Stabil condition is satisfied with $c_0 = 3$ and $\epsilon = \frac{1}{2n}$. Let $\zeta^* = 2\sum_{g=1}^{p} I(\alpha^*_g \neq 0)$. Then, for sufficiently large λ_n and with high probability, we have*

$$\sum_{g=1}^{p} \left\| \left(\hat{\beta}_g - \beta^*_g\right) I(\alpha^*_g \neq 0) \right\|_2 \leq \frac{\max\limits_{g\in\{1,\ldots,p\}}\{|v_g|\}}{\sqrt{2}} \left(\frac{4}{c_n k}\lambda_n \zeta^* + \left(1 + \frac{1}{\lambda_n}\right)\frac{1}{2n} \right)$$

where $0 < k < 1$ and $c_n > 0$ are defined in Chapter 2.8.2.

$\beta^* = (\alpha^*, \gamma^*)^T$ denotes the true/least false coefficient parameters of the outcome and treatment models. The *Group Stabil* condition – a lower bound on the eigenvalues of the covariance matrix – and c_n are discussed in greater detail in Chapter 2.8.2.

The implication of Theorem 1 is that if $\zeta^* = O(1)$, i.e., the number of groups containing non-zero coefficients does not increase with n, then for suitable λ_n (described in Chapter 2.8.2) $\sum_{g=1}^{p} ||(\hat{\beta}_g - \beta^*_g) I(\alpha^*_g \neq 0)||_2 = O\left(\sqrt{\frac{\log p}{n}}\right)$ with high probability, so that $\hat{\alpha}_g \xrightarrow{p} \alpha^*_g$

and $\hat{\gamma}_g \xrightarrow{p} \gamma_g^*$ for all g such that $\alpha_g^* \neq 0$ provided the rate of increase in the number of covariates is $o(e^n)$. Consequently, under the assumptions given above and in Theorem 1, GLiDeR asymptotically recovers all covariates associated with the outcome, which includes the confounders, even when the outcome model is misspecified, and combined with the estimator in (1), yields a consistent estimator of Δ when either the outcome or treatment model is correctly specified. However, we note that GLiDeR is not doubly robust in the fullest sense since the assumptions of Theorem 1 (and above) rule out some particular data generating mechanisms.

2.5 Simulations

2.5.1 Design

We investigate the finite sample behavior of GLiDeR relative to four alternative variable selection approaches to fit models used in treatment effect estimators: (1) the "saturated" method which uses all covariates in fitting the outcome and treatment models to compute $\hat{\Delta}_{DR}$; (2) backward selection on the outcome model (p-stay < 0.05) to select the covariates which are used in fitting the outcome and treatment model to compute $\hat{\Delta}_{DR}$, (3) two-stage model averaged double robust (MADR) estimator proposed by Cefalu et al. (2017), and (4) adaptive lasso (Zou, 2006) to select covariates and estimate the treatment effect using only the outcome model (10-fold cross-validation to select tuning parameter). Note that the first three methods are doubly robust while the adaptive lasso does not consider a model for the treatment. Numerous simulation scenarios are considered to evaluate the effects of varying levels of confounding, model misspecification, covariate structure, number of irrelevant variables (i.e., covariates unrelated to outcome and treatment), and sample size. For each scenario, we generate potential confounders $\mathbf{V} = \{V_1, ..., V_p\}$ marginally as $\mathrm{N}(\mu_v, \sigma_v^2)$, treatment A as Bernoulli[expit$\{f(\mathbf{V})\}$] for some function $f(\cdot)$, where expit(x) $= \frac{\exp(\mathrm{x})}{\exp(\mathrm{x})+1}$, and outcome Y as $\mathrm{N}\{A + g(\mathbf{V}), \sigma_y^2\}$ for some function $g(\cdot)$.

To vary levels of confounding and model misspecification, we consider the same nine distinct combinations of $f(\mathbf{V})$ and $g(\mathbf{V})$ as in Cefalu et al. (2017) (we refer to these as Scenarios 1–9) with both independent and correlated covariates, and one from Ertefaie et al. (2015) (referred to as Scenario 10) with only independent covariates, which are described in Table 2.1. For Scenarios 5–9 (which were used in a previous version of Cefalu et al. (2017))

Table 2.1: Scenarios considered. Treatment A is generated as Bernoulli[expit$\{f(\mathbf{V})\}$], and outcome Y is generated as $N(A + g(\mathbf{V}), \sigma_y^2)$ where $\sigma^2 = 1$ for Scenarios 1–9 and $\sigma_y^2 = 4$ for Scenario 10.

Scenario	$f(\mathbf{V})$ (Treatment)	$g(\mathbf{V})$ (Outcome)
1	$0.4V_1 + 0.3V_2 + 0.2V_3 + 0.1V_4$	0
2	$0.5V_1 + 0.5V_2 + 0.5V_3 + 0.1V_4$	$0.5V_1 + V_3 + 0.5V_4$
3	$0.1V_1 + 0.1V_2 + V_3 + V_4 + V_5$	$2V_1 + 2V_2$
4	$0.5V_1 + 0.4V_2 + 0.3V_3 + 0.2V_4 + 0.1V_5$	$0.5V_1 + V_2 + 1.5V_3 + 2V_4 + 2.5V_5$
5	$0.5V_1 + 0.5V_2 + 0.1V_3$	$V_3 + V_4 + V_5 + \sum_{i=1}^{5}\sum_{j=1}^{5} V_iV_j$
6	$V_1 + V_2 + V_5$	$\sum_{i=1}^{5}\sum_{j=1}^{5} 0.5V_iV_j$
7	$0.2V_1 + 0.2V_2 + 0.2V_5$	$0.25V_3 + (V_1 + V_2)^2 - (V_1^2 - V_3)^2 +$ $(V_4^2 - 0.5V_5)(V_3 - 0.5V_4)$
8	$V_3 + V_4 + V_5 + \sum_{i=1}^{5}\sum_{j=1}^{5} V_iV_j$	$0.5V_1 + 0.5V_2 + 0.1V_3$
9	$(X_1 + X_2 + 0.5X_3)^2$	$0.5V_1 + 0.5V_3 + 0.5V_4$
10	$0.2V_1 - 2V_2 + V_5 - V_6 + V_7 - V_8$	$2V_1 + 0.2V_2 + 5V_3 + 5V_4$

$f(\mathbf{V})$ or $g(\mathbf{V})$ (but not both) is a polynomial function of the covariates, while all methods assume $f(\mathbf{V})$ and $g(\mathbf{V})$ to be linear functions of the covariates, so that the outcome or treatment model (but not both) is misspecified in these scenarios. We also varied the total number of covariates available for Scenarios 1–9 by considering $p = 5, 10$, and 25 (we do not consider MADR for $p = 25$ as this would require fitting over 6 million models for each dataset) with a sample size of $n = 500$, and we varied the sample size by considering $n = 250$ and $n = 500$ with 10 covariates. For Scenario 10 we consider $p = 100$, $p = 500$, and $p = 1000$ with a sample size of $n = 500$ and only consider GLiDeR and the adaptive lasso, and compare them to the saturated method with a ridge penalty for both models due to the large number of covariates. Bootstrap 95% percentile confidence intervals of the treatment effect estimate using GLiDeR are calculated for Scenarios 1–9 and Scenario 10 with $p = 100$ using 1,000 bootstrap samples. All results represent averages over 1,000 Monte Carlo datasets.

2.5.2 Results

Table 2.2 shows the ratio of mean squared error (MSE) of the average causal treatment effect of GLiDeR, backward selection, MADR, and adaptive lasso (denominator) relative to the saturated variable selection method (numerator) and Monte Carlo (MC) bias and

Table 2.2: Ratio of MSE (saturated model MSE / alternative method MSE) and Monte Carlo (MC) bias and standard errors for each scenario with sample size n = 500 over 1,000 MC datasets. Scenarios 1–9 have 10 covariates and Scenario 10 has varying covariate set sizes (p). In the simulations with correlated covariates, $\rho(V_i, V_j) = 0.6$ for $i \neq j \leq 5$ and $\rho(V_i, V_j) = 0$ for $i \neq j > 5$.

Scenario	GLiDeR			Backward selection			MADR			Adaptive lasso		
	MSE Ratio	MC Bias	MC SD	MSE Ratio	MC Bias	MC SD	MSE Ratio	MC Bias	MC SD	MSE Ratio	MC Bias	MC SD
Independent covariates												
1	**1.10**	0.00	0.09	1.02	0.00	0.09	**1.11**	0.00	0.09	**1.12**	0.00	0.09
2	**1.09**	0.00	0.09	1.02	0.00	0.10	**1.12**	0.00	0.09	**1.13**	0.00	0.09
3	**2.91**	0.00	0.09	1.04	0.00	0.15	**3.06**	0.00	0.09	**3.20**	0.00	0.09
4	**1.01**	0.00	0.10	1.01	0.00	0.10	**1.02**	0.00	0.10	**0.79**	0.05	0.10
5	**1.67**	-0.04	0.63	**1.05**	-0.04	0.80	**1.65**	-0.03	0.64	**1.64**	-0.04	0.64
6	**18.35**	0.00	0.31	0.95	-0.03	1.36	**18.53**	0.00	0.31	**16.30**	0.00	0.33
7	**1.14**	-0.05	0.84	**1.04**	-0.06	0.88	**1.12**	-0.05	0.85	**1.16**	-0.05	0.83
8	**1.26**	0.01	0.12	**1.04**	0.01	0.13	**1.25**	0.02	0.12	**1.35**	0.01	0.11
9	**1.05**	0.00	0.11	**1.04**	0.00	0.11	**1.06**	0.00	0.11	**1.06**	0.00	0.11
Correlated covariates												
1	**1.16**	0.00	0.09	1.01	0.00	0.10	**1.20**	0.00	0.09	**1.20**	0.00	0.09
2	**1.09**	0.00	0.10	1.02	0.00	0.10	**1.09**	0.00	0.10	**1.14**	0.02	0.10
3	**5.55**	0.00	0.13	**0.94**	-0.02	0.32	**5.29**	-0.01	0.14	**7.49**	0.02	0.11
4	**1.05**	0.01	0.11	**1.04**	0.01	0.11	**1.08**	0.01	0.11	**0.25**	0.20	0.11
5	**3.30**	0.39	1.99	1.02	0.05	3.64	**2.68**	0.45	2.20	**2.14**	0.99	2.15
6	**305.39**	-0.02	0.92	**1.61**	-0.10	12.69	**74.41**	0.02	1.86	**199.97**	0.05	1.14
7	**1.27**	0.06	0.86	**1.09**	0.05	0.93	**1.15**	0.07	0.91	**1.96**	0.05	0.69
8	**1.28**	0.01	0.15	**1.12**	0.01	0.16	**1.29**	0.02	0.15	**1.29**	0.02	0.15
9	**1.05**	0.00	0.11	**1.06**	0.00	0.11	**1.06**	0.00	0.11	**1.06**	0.00	0.11
Scenario 10												
$p = 100$	**1.60**	0.00	0.26	*	*	*	*	*	*	**0.51**	-0.42	0.20
$p = 500$	**13.65**	-0.06	0.32	*	*	*	*	*	*	**6.22**	-0.42	0.20
$p = 1000$	**13.76**	-0.14	0.29	*	*	*	*	*	*	**7.43**	-0.39	0.20

Bold indicates significant difference (5% significance level) between MSEs (testing equality) from the saturated method (full model) vs. the alternative method using the paired t-test.

standard deviation for a sample size of 500 and 10 covariates (Scenarios 1-9). Additional results for different sample sizes and number of covariates are given in Chapter 2.8.3. Note that a larger value for the MSE ratio indicates better performance, with a MSE ratio greater than one demonstrating improved treatment effect estimation over the saturated method.

Apart from a few exceptions, all MSE ratios are greater than one as including all covariates available (saturated method) generally led to treatment effect estimates with higher MC variance. Additionally, backward selection shows a smaller MSE ratio than MADR and GLiDeR in all scenarios except one (Scenario 9 with correlated covariates, where the three ratios are similar) as using backward selection on the outcome model to select covariates for treatment effect estimation was generally more variable than performing variable selection across both the outcome and treatment models (GLiDeR, MADR).

Comparing GLiDeR and MADR, when both models were specified correctly (Scenarios 1–4) or when only the treatment models were misspecified (Scenarios 8 and 9), GLiDeR

and MADR performed similarly. However, in all scenarios where the outcome models were misspecified (Scenarios 5–7) with correlated covariates, GLiDeR displayed a less variable treatment effect estimator and significantly greater MSE ratio than MADR; the methods performed similarly in these scenarios with independent data.

The adaptive lasso approach considered here uses only the outcome model for estimation of the treatment effect and is, therefore, more efficient than doubly robust methods when the outcome model is correctly specified. This approach outperformed all methods with a correctly specified outcome model, except Scenarios 4 and 10 with $p = 100$, where it displayed much greater MC bias and performed significantly worse (MSE ratio < 1) than all methods. In these two scenarios, there is a confounder weakly associated with the outcome but strongly related to treatment (V_1 in Scenario 4 and V_2 in Scenario 10). The adaptive lasso tends to omit these variables as it considers only the associations in the outcome model and ignores the relationships between treatment and covariates. Excluding these important confounders in Scenarios 4 and 10 introduces a large bias and consequently larger MSE compared to the other methods. GLiDeR, however, selects this important confounder in nearly all datasets in these scenarios (see Table 2.8 for percentage of datasets each covariate is selected by GLiDeR) and accordingly has much smaller bias than adaptive lasso using only the outcome model. In Scenario 10 when the number of irrelevant covariates is increased ($p = 500$ and $p = 1000$; $n = 500$), the bias of GLiDeR also increases as it becomes more challenging to select X_2, but the bias is much smaller than that of adaptive lasso and, even with a larger variance, GLiDeR has an MSE ratio approximately twice that of the adaptive lasso. Even with the large bias of the adaptive lasso, it is more efficient than using all covariates with ridge penalty with $p = 500$ and $p = 1000$.

GLiDeR achieved coverage rates very close to the nominal 95% in all scenarios that were considered for confidence interval coverage (see Table 2.9).

2.5.3 Computation time

GLiDeR is dramatically faster than MADR, making it feasible to apply in problems where p is much larger. With $p = 10$ covariates and sample size $n = 500$, GLiDeR required 3 seconds while MADR required 10. However, the computation time of GLiDeR scales linearly with the number of covariates p, while the computation time of MADR scales exponentially as 2^p. For instance, MADR would take over $1,000$ hours with 30 covariates

(ignoring the time for storage and other necessary calculations), while GLiDeR solves the same size problem in less than 20 seconds; in Scenario 10 with sample size $n = 500$ and $p = 100$, $p = 500$, and $p = 1000$, GLiDeR took approximately 20 seconds, 3.5 minutes, and 11 minutes, respectively, to compute the ACE per dataset over a sequence of 100 λ. All computations were performed using a pure `R` implementation, rather than a faster language like C.

2.6 Application

Bilateral lung transplant (BLT) is generally associated with lower short-term survival, but higher quality of life compared to single-lung transplant (SLT) for individuals with lung disease (Aziz et al., 2010). Consequently, the effect of BLT (vs. SLT) on physiologic measures associated with quality of life, such as forced expiratory volume in one second (FEV1), is important for patients who must decide between the two treatment options. Data on lung transplant recipients from May 2005 – September 2011 were obtained from the United Network for Organ Sharing national registry. In this analyses, we focus on patients aged 60 or older with obstructive lung disease (e.g., COPD). The dataset consists of 937 patients (52.7% receiving BLT) and 31 potential confounders, which are summarized in Table 2.10. Missing covariate data were imputed using Multivariate Imputation by Chained Equations (MICE) (Van Buuren and Groothuis-Oudshoorn, 2011). The outcome is FEV1% one year after transplant, where FEV1% is defined as the percentage of the predicted value of FEV1 given the person's age, height, gender, and race. Patients who died were given an FEV1 of 0, the worst possible score. We assume linear and logistic regression models for the outcome and treatment (BLT vs. SLT), respectively. With 31 covariates, we would have to fit 2^{32} (over 4 billion) models to estimate the treatment effect using model averaged methods, making GLiDeR an appealing option.

We estimated coefficient values for a sequence of 100 λ values ranging from 0 to the smallest value of λ such that all coefficients are zero. We then selected the optimal λ (denoted λ^*) using GCV on the outcome model as described in Chapter 2.3.4. For all methods, 1,000 bootstrap samples were used to estimate standard errors and obtain 95% percentile-based confidence intervals (CIs) of the treatment effect estimates. Figure 2.1 shows the estimated coefficients from the outcome and treatment models for a subset of

Table 2.3: Variables selected and estimated coefficients (for standardized variables and outcome) by GLiDeR and backward selection.

	GLiDeR		Backward selection	
Covariate	Outcome Coef	Treatment Coef	Outcome Coef	Treatment Coef
Ischemic time	-0.075	-1.018	-0.060	-1.154
Age of recipient	0.097	0.171	0.114	0.270
PO2	0.032	-0.014	0.060	-0.079
Oxygen amount required	-0.052	-0.088	-0.060	-0.261
6 minute walk distance	0.019	-0.004	0.061	-0.044
Height of recipient	-0.058	0.008	*	*
Height of donor	-0.015	0.005	*	
Local or regional (vs. national) allocation	0.034	0.096	*	*
Center volume	0.010	-0.013	*	*
Sex of recipient	*	*	0.092	-0.034

*Covariate was not chosen by method

λ values that were considered. Table 2.10 displays the selected covariates and estimated coefficients by GLiDeR and backward selection; nine covariates were selected by GLiDeR and six covariates were chosen by backward selection for final estimation of the treatment effect. Figure 2.2 displays a forest plot comparing point estimates and 95% CIs of the ACE of BLT (vs. SLT) on FEV1% one year after transplant for GLiDeR, backward selection, and the saturated method.

Using backward selection on the outcome model as described in Chapter 2.5.1, the ACE is estimated to be 34.7 with a standard error of 3.4, both equivalent (to one decimal place) to the estimates obtained using all covariates, but with a slightly smaller 95% CI: (27.4, 38.7) with backward selection compared to (26.1, 39.1) using all covariates. The standard errors and CI length using these methods are much larger than those achieved with GLiDeR, where the estimated coefficients at λ^* are used in the standard doubly robust estimator in Equation (2.1) and the ACE of BLT (vs. SLT) is estimated to be 36.0 (FEV1% after one year) with a corresponding standard error of 1.6 and 95% CI of (32.6, 38.9).

These results are consistent with simulations, where GLiDeR generally shows greater efficiency over these methods as the number of covariates is increased (see Tables 2.4 and 2.6). Even though the differences in estimated treatment effects between GLiDeR and other approaches appear small, the difference in sample means of FEV1% among the treated (BLT) and untreated (SLT) is 33.9, meaning the gap between the effect estimate from GLiDeR and other methods which incorporate covariates is larger than that between those other methods and the sample mean difference. In settings where incorporating covariates makes a bigger difference to the treatment effect estimate, GLiDeR may offer a substantial gain in efficiency.

Figure 2.1: Coefficient estimates for the outcome (top) and treatment (bottom) models. A white box indicates a coefficient is equal to zero, while a darker box indicates a coefficient is larger in magnitude. Variables are ordered by the magnitude of their outcome model coefficients at $\lambda = 0$ (unpenalized model) from largest to smallest.

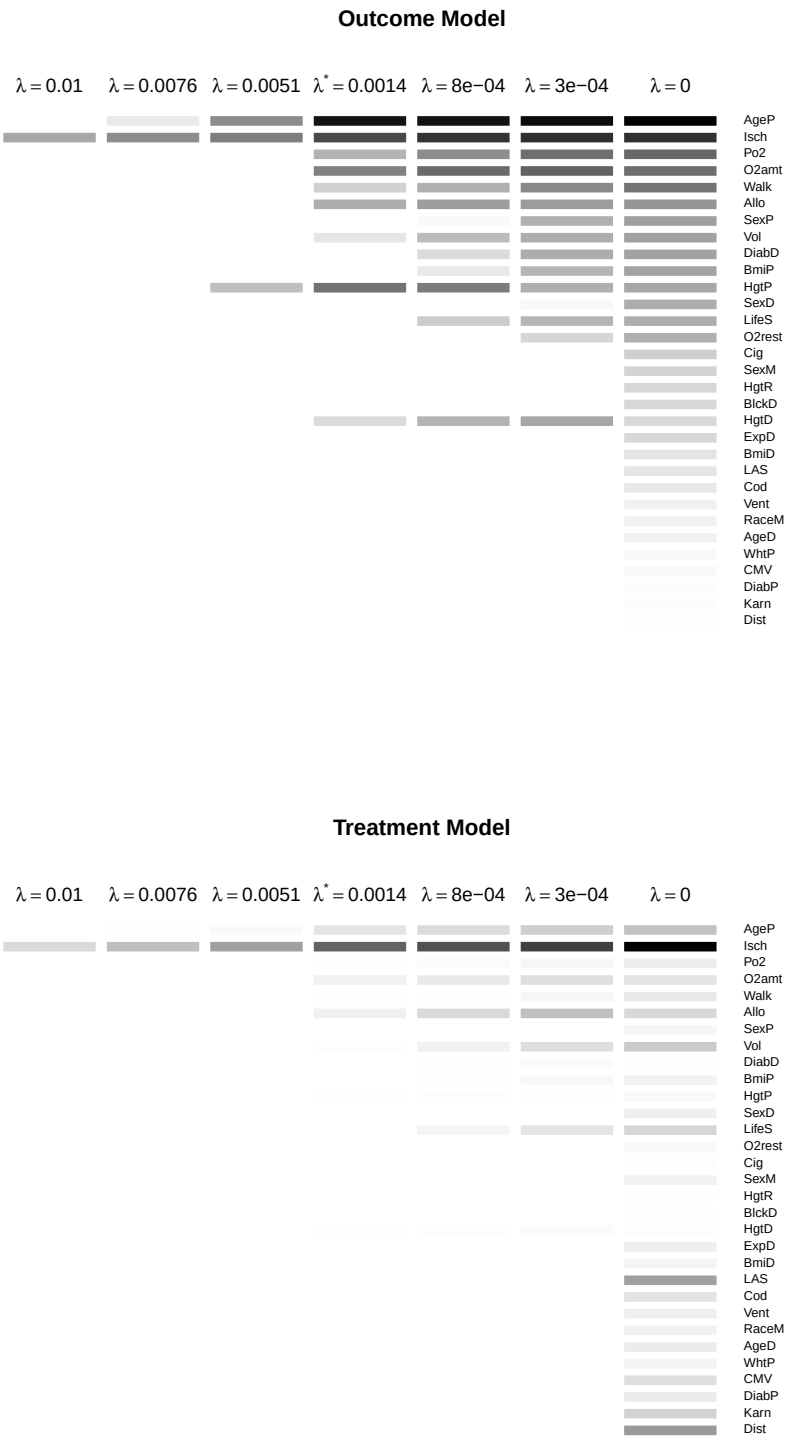
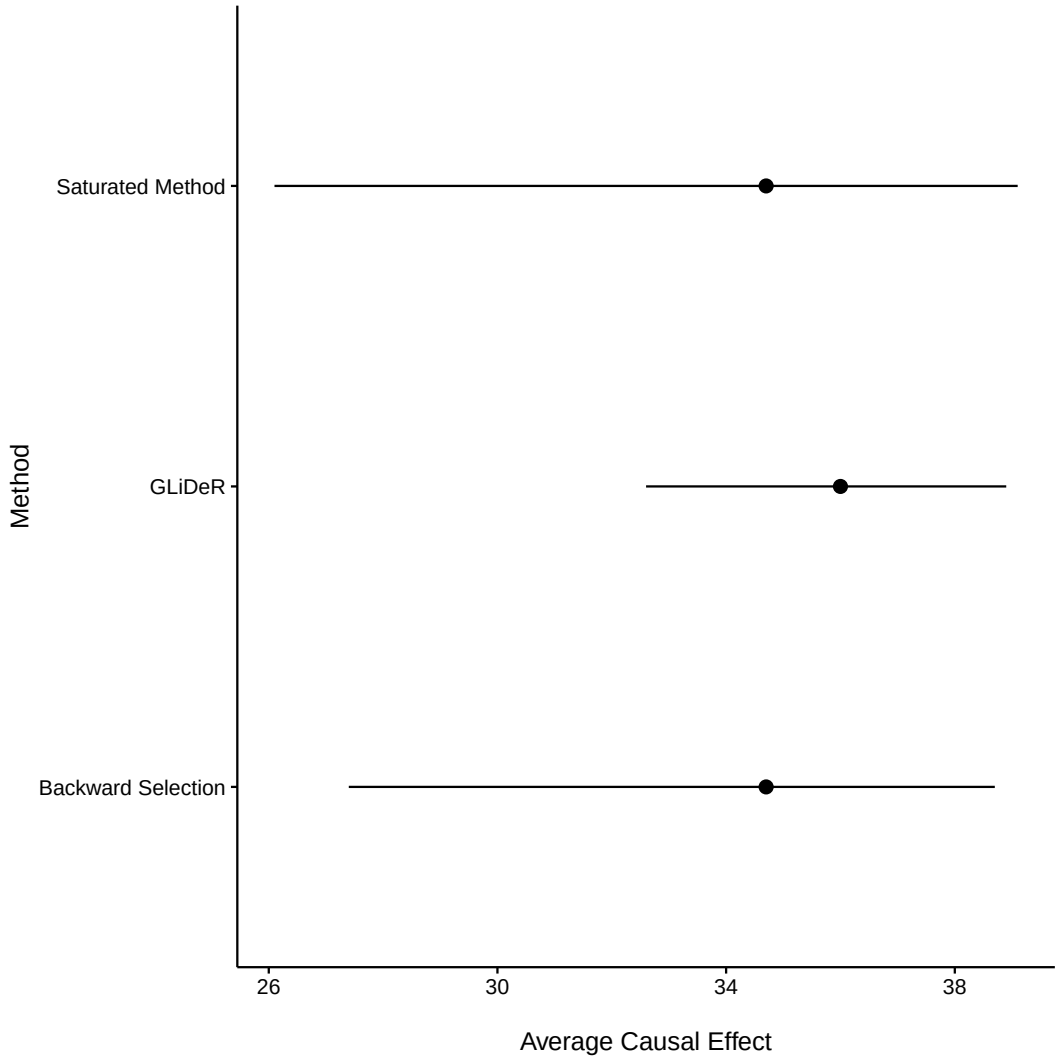


Figure 2.2: Forest plot of point estimates and corresponding Bootstrap percentile 95% confidence intervals of the ACE of BLT (vs. SLT) on FEV1% one year after transplant for GLiDeR, backward selection, and the saturated method.



2.7 Discussion

Doubly robust estimation of the average causal treatment effect requires working models for both the outcome and treatment given possible confounders. When the number of possible confounders is large it is natural to consider some form of variable selection for the outcome and treatment models. GLiDeR uses an adaptive group lasso approach to perform coefficient regularization and estimation across both treatment and outcome models simultaneously, unlike traditional methods that consider only one model and are thus more likely to exclude important confounders with weak associations in the model under consideration. GLiDeR has desirable theoretical properties, and in simulation experiments outperforms doubly robust approaches which do not incorporate variable selection. It achieves similar efficiency with existing techniques which perform variable selection across both outcome and treatment models, but has substantial computational advantages over these approaches and allows for situations with $p > n$. Simulations suggest the largest gains in efficiency are achieved when the outcome is misspecified, a frequent occurrence in practice.

GLiDeR targets inference for the average causal treatment effect, Δ. Even though GLiDeR displays good performance in the simulation scenarios considered in this chapter, we caution that, like other model selection procedures, its finite sample performance at certain local alternatives can potentially be quite poor, reminiscent of Hodges' estimator (Leeb and Pötscher, 2008). While the validity of bootstrap intervals was not explored in this chapter, percentile bootstrap confidence intervals for Δ had good coverage; how to adapt promising recent developments in post-selection inference to our setting is an area of future research.

2.8 Supplementary Materials

2.8.1 Supplement A: Existence of Unique Solution

Here, we present conditions under which the simultaneous variable selection problem defined by Equation (2.3) in Chapter 2.3.2 has a unique solution. An immediate corollary is that a solution exists when Φ_{sum} is given by a sum of the squared error and logistic loss, i.e., when defining linear and logistic regression models for the outcome and treatment, respectively.

For notational purposes, let $\mathbf{D}$ denote the working data $\{Y, A, \mathbf{V}\}$ and $L(\beta|\mathbf{D})$ be the

empirical loss, i.e.,

$$L(\beta|\mathbf{D}) = \frac{1}{n}\sum_{i=1}^{n}\Phi_{sum}(Y_i, A_i, \mathbf{V}_i; \beta). \tag{2.6}$$

Yang and Zou (2015) show that Equation (2.3) has a solution provided the loss function Φ_{sum} satisfies a so-called quadratic majorization (QM) condition, i.e., if and only if the following two assumptions hold:

(i) $L(\beta|\mathbf{D})$ is a differentiable function of β, i.e., $\nabla L(\beta, \mathbf{D})$ exists everywhere.

(ii) There exists a p x p matrix $\mathbf{H}$, which may only depend on $\mathbf{D}$, such that for all β, β^*

$$L(\beta|\mathbf{D}) \leq L(\beta^*|\mathbf{D}) + (\beta - \beta^*)^T\nabla L(\beta^*|\mathbf{D}) + \frac{1}{2}(\beta - \beta^*)^T\mathbf{H}(\beta - \beta^*)$$

We state and prove the following extension to their result which characterizes a class of loss functions of the form of Equation (2.2) that satisfy the QM condition:

Lemma 1. *Let $\Phi_{sum}(Y, A, f, g) = \Phi_{out}(Y, f) + \Phi_{trt}(A, g)$, where Φ_{out} is the loss function used to link outcome Y with predictors $Z_1 = \{Z_{11}, \ldots, Z_{1r}\}$ through a linear predictor $f = \alpha^T Z_1$, and Φ_{trt} is the loss function used to link treatment A with predictors $Z_2 = \{Z_{21}, \ldots, Z_{2s}\}$ through a linear predictor $g = \gamma^T Z_2$. Let $Z = \{Z_{11}, \ldots, Z_{1r}, Z_{21}, \ldots, Z_{2s}\}$. Assume Φ_{out} is differentiable with respect to the coefficient parameters in f and write $\Phi'_{out} = \frac{\partial\Phi_{out}(Y,f)}{\partial f}$, and similarly, assume Φ_{trt} is differentiable with respect to the coefficient parameters in g and write $\Phi'_{trt} = \frac{\partial\Phi_{trt}(A,g)}{\partial g}$. Then:*

(1). If Φ'_{out} and Φ'_{trt} are Lipschitz continuous with constants C_1 and C_2 such that

$$(i)\ |\Phi'_{out}(Y, f_1) - \Phi'_{out}(Y, f_2)| \leq C_1|f_1 - f_2|\ \forall Y, f_1, f_2,$$

and

$$(ii)\ |\Phi'_{trt}(A, g_1) - \Phi'_{trt}(A, g_2)| \leq C_2|g_1 - g_2|\ \forall A, g_1, g_2,$$

then the QM condition holds for Φ_{sum} and $\boldsymbol{H} = \frac{2(C_1+C_2)}{n}\mathbf{Z}^T\mathbf{Z}$.

(2). If $\Phi''_1 = \frac{\partial\Phi^2(Y,f)}{\partial f^2}$ and $\Phi''_2 = \frac{\partial\Phi^2(A,g)}{\partial g^2}$ exist and there are constants C_3 and C_4 such that

$$(i)\ \Phi''_1 \leq C_3\ \forall Y, f,$$

and

$$(ii)\ \Phi_2'' \leq C_4\ \forall A, g$$

then the QM condition holds for Φ_{sum} *and* $\boldsymbol{H} = \frac{C_3+C_4}{n}\mathbf{Z}^T\mathbf{Z}$.

(3) If

(i) Φ_{out} *satisfies condition (1)(i) with constant* C_1,

(ii) Φ_{trt} *satisfies condition (2)(ii) with constant* C_2,

and

(iii) $\Phi_2'' = \frac{\partial\Phi^2(A,g)}{\partial g^2} \geq C_L\ \forall A, g$ *(i.e.,* Φ_2'' *is bounded),*

or

(i) Φ_{trt} *satisfies condition (1)(ii) with constant* C_1,

(ii) Φ_{out} *satisfies condition (2)(i) with constant* C_2,

and

(iii) $\Phi_1'' = \frac{\partial\Phi^2(Y,f)}{\partial f^2} \geq C_L\ \forall Y, f$ *(i.e.,* Φ_1'' *is bounded),*

then the QM condition holds for Φ_{sum} *and* $\boldsymbol{H} = \frac{2(C_1+C_2^*)}{n}\mathbf{Z}^T\mathbf{Z}$, *where* $C_2^* = \max\{|C_2|, |C_L|\}$.

Proof. Before proving Lemma 1, we first present a lemma (without observation weights) from Yang and Zou (2015):

Lemma 2. *Assume* $\Phi(y, f)$ *is differentiable with respect to* f *and write* $\Phi_f' = \frac{\partial\Phi(y,f)}{\partial f}$. *Then*

$$\nabla\Phi(\beta|\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n}\tau_i\Phi'(y_i, x_i^T\beta)x_i$$

(1). If Φ_f' *is Lipschitz continuous with constant* C *such that*

$$|\Phi_f'(y, f_1) - \Phi_f'(y, f_2)| \leq C|f_1 - f_2|\ \forall y, f_1, f_2,$$

then the QM condition holds for Φ *and* $\boldsymbol{H} = \frac{2C}{n}\mathbf{X}^T\mathbf{X}$.

(2). If $\Phi_f'' = \frac{\partial\Phi^2(y,f)}{\partial f^2}$ *exists and* $\Phi_f'' \leq C_2\ \forall y, f$,

then the QM condition holds for Φ *and* $\boldsymbol{H} = \frac{C_2}{n}\mathbf{X}^T\mathbf{X}$.

Proving (1): We have

$$|\Phi'_{out}(Y, f_1) - \Phi'_{out}(Y, f_2)| \leq C_1|f_1 - f_2| \; \forall Y, f_1, f_2,$$

and

$$|\Phi'_{trt}(A, g_1) - \Phi'_{trt}(A, g_2)| \leq C_2|g_1 - g_2| \; \forall A, g_1, g_2$$

Then condition (1) in Lemma 2 is satisfied for the outcome and treatment loss functions with constants C_1 and C_2, respectively. This implies

$$|\Phi'_{sum}(Y, A, f_1, f_1) - \Phi'_{sum}(Y, A, f_2, f_2)|$$

$$= |\Phi'_{out}(Y, f_1) + \Phi'_{trt}(A, f_1) - \Phi'_{out}(Y, f_2) - \Phi'_{trt}(A, f_2)|$$

$$\leq |\Phi'_{out}(Y, f_1) - \Phi'_{out}(Y, f_2)| + |\Phi'_{trt}(A, f_1) - \Phi'_{trt}(A, f_2)|$$

$$\leq C_1|f_1 - f_2| + C_2|f_1 - f_2| \; \forall Y, A, f_1, f_2.$$

To prove (2) in Lemma 1: We have constants C_3 and C_4 such that $\Phi''_{out}(Y, f) \leq C_3$ and $\Phi''_{trt}(A, g) \leq C_4$ for all Y, f, g. Then

$$\Phi''_{sum}(Y, A, f, g) = \Phi''_{out}(Y, f) + \Phi''_{trt}(A, g) \leq C_3 + C_4.$$

Finally, to prove (3) in Lemma 1: Assume condition (1) in Lemma 2 is satisfied for, say (WLOG), Φ_{out} with constant C_1, and also assume Φ_{trt} satisfies condition (2) in Lemma 2 with constant C_2 such that $\Phi''_{trt} \geq C_L$. Then since Φ''_{trt} is bounded, we know Φ'_{trt} is Lipschitz continuous with constant C_2^* (bounded derivative implies Lipschitz continuity). The proof then concludes following the proof of (1) with constants C_1 and C_2^*. □

To use linear and logistic regression to model the outcome and treatment, respectively, and (naturally) letting Φ_{out} be the squared-error loss function and Φ_{trt} be the loss function proportional to the binomial log-likelihood, we have $\Phi''_{ls} = 1$ and $\Phi''_{logit} \leq \frac{1}{4}$, meaning the QM condition holds for $\mathbf{H} = \frac{(5/4)}{n}\mathbf{Z}^T\mathbf{Z}$ by Lemma 1 condition (2).

When the QM condition is met (i.e., when the conditions of Lemma 1 are satisfied), we are able to solve for β in Equation (2.3) using the groupwise-majorization-descent (GMD) algorithm (for details, see Yang and Zou (2015)), a computationally efficient and unified

algorithm allowing for general design matrices.

2.8.2 Supplement B: Proof of Theorem 1

Here, we provide a proof of Theorem 2 in the main text, which is re-stated here:

Theorem 2. *Assume the number of covariates p and sample size n are such that $\frac{\log(2p)}{n} \leq 1$. Also assume the Group Stabil condition is satisfied with $c_0 = 3$ and $\epsilon = \frac{1}{2n}$. Let $\zeta^* = 2\sum_{g=1}^{p} I(\alpha_g^* \neq 0)$. Then, for sufficiently large λ_n and with high probability, we have*

$$\sum_{g=1}^{p} \left\| \left(\hat{\beta}_g - \beta_g^*\right) I(\alpha_g^* \neq 0) \right\|_2 \leq \frac{\underset{g\in\{1,\ldots,p\}}{max}\{|v_g|\}}{\sqrt{2}} \left(\frac{4}{c_n k} \lambda_n \zeta^* + \left(1 + \frac{1}{\lambda_n}\right) \frac{1}{2n} \right)$$

where $0 < k < 1$ is defined in Definition 1 (pg. 12), and

$$c_n = \min_{[|x|<L(9B+\frac{1}{n})]\cap\Theta} \left\{ \frac{\Psi''_{out}(x) + \Psi''_{trt}(x)}{2} \right\}.$$

We recall that λ_n is the penalty, $\hat{\beta}$ is the group lasso estimator in our set-up, β^* is the vector of true/least false coefficient parameters in the outcome and treatment models, and β_g^* is the sub-vector of β^* associated with group g (in our case, covariate g). We let p^* be the total number of columns in the design matrices of the outcome and treatment models (i.e, p^* is the length of β^*), which we denote by $\mathbf{Z}_{out}$ and $\mathbf{Z}_{trt}$, respectively. We assume ($Z_{out,i}$,$Z_{trt,i}$,Y_i,A_i) are i.i.d. copies of (Z_{out},Z_{trt},Y,A) for $i = 1, \ldots, n$, where $Y|Z_{out}$ and $A|Z_{trt}$ are modeled by distributions F_{out} and F_{trt} both on $\mathbb{R}$ and from the exponential family, respectively, and $Z_{out,i}$ and $Z_{trt,i}$ are the ith rows of $\mathbf{Z}_{out}$ and $\mathbf{Z}_{trt}$, respectively. The natural parameter space is denoted by $\Theta := \Theta_{out} \cup \Theta_{trt}$, where

$$\Theta_{out} = \left\{ \theta \in \mathbb{R} : \int \exp(\theta x) F_{out}(dx) < \infty \right\}$$

and

$$\Theta_{trt} = \left\{ \theta \in \mathbb{R} : \int \exp(\theta x) F_{trt}(dx) < \infty \right\}.$$

L and B apply to assumptions (H.1–3):

(H.1): the pair of variables (Z_{out}, Z_{trt}) are almost surely bounded by a constant L, i.e., there exists a constant $L > 0$ such that

$||(Z_{out}, Z_{trt})||_\infty \leq L$ a.s.

(H.2): for all $x \in [-L, L]^{p^*}$, $\beta^{*T} x \in \text{Int}(\Theta)$

(H.3): There exists a constant $B > 0$ such that $\sum_{g=1}^{G_n} \sqrt{d_g} ||\beta_g^*||_2 \leq B$

We consider $\Lambda = \left\{\beta \in \mathbb{R}^{p^*} : \forall x \in [-L, L]^{p^*}, \beta^T x \in \Theta\right\}$. We define d_g to be the size of group g, $g \in \{1, \ldots, G_n\}$, and let $d_{min} := \min_{g \in \{1,\ldots,G_n\}} d_g$ and $d_{max} := \max_{g \in \{1,\ldots,G_n\}} d_g$ denote the smallest and largest group sizes, respectively. Letting $\Phi_{sum}(\beta) = \Phi_{sum}(Y, A, \mathbf{Z}; \beta)$, the empirical process $(\mathbb{P}_n - \mathbb{P})(\Phi_{sum}(\beta))$ can be written as:

$$\begin{aligned}(\mathbb{P}_n - \mathbb{P})(\Phi_{sum}(\beta)) &= (\mathbb{P}_n - \mathbb{P})[\Phi_{out}(\beta) + \Phi_{trt}(\beta)] \\ &= (\mathbb{P}_n - \mathbb{P})[\Phi_{out,l}(\beta)] + (\mathbb{P}_n - \mathbb{P})[\Phi_{out,\Psi}(\beta)] \\ &+ (\mathbb{P}_n - \mathbb{P})[\Phi_{trt,l}(\beta)] + (\mathbb{P}_n - \mathbb{P})[\Phi_{trt,\Psi}(\beta)],\end{aligned}$$

where $\Phi_{out,l} = -Y\alpha'\mathbf{Z}_{out}$, $\Phi_{out,\Psi} = \Psi_{out}(\alpha'\mathbf{Z}_{out})$, $\Phi_{trt,l} = -A\gamma'\mathbf{Z}_{trt}$, and $\Phi_{trt,\Psi} = \Psi_{trt}(\gamma'\mathbf{Z}_{trt})$; $\Psi''_{out}(x)$ and $\Psi''_{trt}(x)$ in Theorem 2 denote the second derivatives of $\Psi_{out}(x)$ and $\Psi_{trt}(x)$, respectively. $\Phi_{\cdot,l}$ is used to denote the *linear* part of Φ and $\Phi_{\cdot,\Psi}$ is used to denote the part which depends on the link function between the canonical parameter and the linear predictor. For example, if modeling the outcome with linear regression (i.e., using squared error loss), then $\Phi_{out,\Psi} = \alpha'\mathbf{Z}_{out}$, and if modeling the treatment with logistic regression, $\Phi_{trt,\Psi} \propto n \log(1 + \exp(\gamma'\mathbf{Z}_{trt}))$.

We define

$$L_g := \left\| \frac{\hat{\beta}_g^{ls}}{\sqrt{d_g}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i}^g \right)^T - E\left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt}^g \right)^T \right\} \right\|_2$$

for all $g \in \{1, \ldots, G_n\}$, where $Z_{out,i}^g$ and $Z_{trt,i}^g$ denote the elements on the ith rows and gth columns of $\mathbf{Z}_{out}$ and $\mathbf{Z}_{trt}$, respectively. We also define

$$\mathcal{A} = \bigcap_{g=1}^{G_n} \left\{ L_g \leq \frac{\lambda_n}{2} \right\}$$

and

$$\mathcal{B} = \left\{ \sup_{\beta : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g - \beta_g^*||_2 \leq M} |v_n(\beta, \beta^*)| \leq \frac{\lambda_n}{2} \right\}$$

where

$$v_n = \frac{(\mathbb{P}_n - \mathbb{P})\left([\Phi_{out,\Psi}(\beta^*) - \Phi_{out,\Psi}(\beta)] + [\Phi_{trt,\Psi}(\beta^*) - \Phi_{trt,\Psi}(\beta)]\right)}{\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g - \beta_g^*||_2 + \epsilon_n}$$

with $M = 8\left(\min_{g \in G_n}\left\{\hat{\beta}_g^{ls}\right\}\right) B + \epsilon_n$ and $\epsilon_n = \frac{1}{n}$.

We can then adapt the following propositions of Blazère et al. (2014):

Proposition 1. *Provided the penalty term λ_n is chosen suitably large enough,*

$$P(\mathcal{A} \cap \mathcal{B}) \geq 1 - \frac{2(C+2)}{(2G_n)^{A^2/2}}$$

for any $A > \sqrt{2}$, where C is a universal constant.

(Proof at the end of this subchapter) In other words, for some suitable values of λ_n and provided $G_n \to \infty$, the event $\mathcal{A} \cap \mathcal{B}$ happens with probability tending to one, implying the events $\mathcal{A}$ and $\mathcal{B}$ each also have probability tending to one. Propositions 2 and 3 below provide upper bounds for the linear and non-linear parts of the empirical process on the events $\mathcal{A}$ and $\mathcal{A} \cap \mathcal{B}$, each occurring with high probability (by Proposition 1), respectively:

Proposition 2. *On the event $\mathcal{A}$,*

$$(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,l}(\beta^*) - \Phi_{sum,l}(\hat{\beta})) \leq \frac{\lambda_n}{2} \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2.$$

Proof. We have

$$(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,l}(\beta^*) - \Phi_{sum,l}(\hat{\beta}))$$

$$= \sum_{g=1}^{G_n} (\hat{\beta}_g - \beta_g^*)^T \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i}^g\right)^T - E\left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt}^g\right)^T\right]$$

$$\leq \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} \left\|\hat{\beta}_g - \beta_g^*\right\|_2 \left\| \frac{\hat{\beta}_g^{ls}}{\sqrt{d_g}} \frac{1}{n} \sum_{i=1}^{n}\left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i}^g\right)^T - E\left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt}^g\right)^T \right\|_2.$$

The last line follows from the Cauchy-Schwarz inequality, and the proposition follows on the event $\mathcal{A}$. □

Lemma 3. *On the event $\mathcal{A} \cap \mathcal{B}$ we have $\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 \leq M$, where we recall that $M = 8\left(\min_{g \in G_n}\left\{\hat{\beta}_g^{ls}\right\}\right) B + \epsilon_n$ with $\epsilon_n = \frac{1}{n}$.*

Lemma 3 bounds the difference between the estimated and true coefficients and is proved at the end of this subchapter. The next proposition provides an upper bound for $(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\psi}(\beta^*) - \Phi_{sum,\psi}(\hat{\beta}))$ and directly results from Lemma 3 and definition of $\mathcal{B}$.

Proposition 3. *On the event $\mathcal{A} \cap \mathcal{B}$,*

$$(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\hat{\beta})) \leq \frac{\lambda_n}{2}\left(\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \epsilon_n\right)$$

Lemma 4. *Assume assumptions (H.1-3) are fulfilled. For all $k \in \mathbb{N}^*$, there exists constants $C_{L,B}^{out}$ and $C_{L,B}^{trt}$ (which both only depend on L and B) such that $E(|Y|^k) \leq k!(C_{L,B}^{out})^k$ and $E(|A|^k) \leq k!(C_{L,B}^{trt})^k$.*

L applies to assumption (H.1) and is a uniform bound for the maximum magnitude of the covariates, and B applies to assumption (H.3) and bounds the l_2 norm of the true (grouped) covariates. Lemma 4 provides moment bounds for outcome Y and treatment A and follows from Lemma 3.2 in Blazère et al. (2014) when $\sqrt{d_g}$ in assumption (H.2) in Blazère et al. (2014) is replaced with $\frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}}$.

Theorem 2 requires that the *Group Stabil Condition* be satisfied. We state it here:

Definition 1. *Let $\Sigma = \mathbb{E}[(Z_{out}, Z_{trt})(Z_{out}, Z_{trt})^T]$. Define $H^* = \{g : \beta_g^* \neq 0\}$, the index set of the groups for which the corresponding sub vectors of β^* are non-zero. Let c_0 and $\epsilon > 0$ be given. Then Σ satisfies the Group Stabil condition if there exists $0 < k < 1$ such that*

$$\delta^T \Sigma \delta \ \geq k \sum_{g \in H^*} ||\delta^g||_2^2 - \epsilon$$

for any $\delta \in S(c_0, \epsilon)$, where $S(c_0, \epsilon)$ is called the restricted set and is defined for c_0 and $\epsilon > 0$ as $S(c_0, \epsilon) = \{\delta : \sum_{g \in H^{*c}} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\delta^g||_2 \leq c_0 \sum_{g \in H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\delta^g||_2 + \epsilon\}$. A Σ which satisfies the Group Stabil Condition is said to be GS(c_0, ϵ, k).

Definition 1 is similar to the Group Stabil Condition proposed in Blazère et al. (2014), the only difference is that $\sqrt{d_g}$ in the restricted set in Blazère et al. (2014) is replaced here by $\frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}}$ in the restricted set, $S(c_0, \epsilon)$. The Group Stabil Condition places a lower bound on the eigenvalues of the variance matrix, with the lower bound depending on the number of non-zero covariate groups. In other words, it restricts the degree of correlation between covariates in the design matrix.

We can now prove Theorem 2 presented in Chapter 2.4.2:

Proof of Theorem 2:

The proof uses arguments similar to those in Blazère et al. (2014). Using the definition of $\hat{\beta}$, where we recall from the main manuscript that

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \mathbb{P}_n(\Phi_{sum}(\beta)) + \lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g||_2 \right\},$$

we have

$$\mathbb{P}_n \Phi_{sum}(\hat{\beta}) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g||_2 \leq \mathbb{P}_n \Phi_{sum}(\beta^*) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g^*||_2. \tag{2.7}$$

Hence we get (adding $\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*))$ to both sides)

$$\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*)) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g||_2$$

$$\leq (\mathbb{P}_n - \mathbb{P})(\Phi_{sum}(\beta^*) - \Phi_{sum}(\hat{\beta})) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g^*||_2. \tag{2.8}$$

From Proposition 2 and 3 and by adding $\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2$ to both sides of the inequality (2.8) we find, on $\mathcal{A} \cap \mathcal{B}$, that

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*))$$

$$\leq 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} (||\hat{\beta}_g - \beta_g^*||_2 + ||\beta_g^*||_2 - ||\hat{\beta}_g||_2) + \frac{\lambda_n}{2} \epsilon_n.$$

If $g \notin H^*$, where we recall from Definition 1 that $H^* = \{g : \beta_g^* \neq 0\}$ (i.e., the index set of the groups for which the corresponding sub vectors of β^* are non-zero), then $||\hat{\beta}_g - \beta_g^*||_2 + ||\beta_g^*||_2 - ||\hat{\beta}_g||_2 = 0$ and otherwise $||\beta_g^*||_2 - ||\hat{\beta}_g||_2 \leq ||\hat{\beta}_g - \beta_g^*||_2$. So the last inequality can be bounded by

$$4\lambda_n \sum_{g \in H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \frac{\lambda_n}{2} \epsilon_n. \tag{2.9}$$

By the definition of β^* we have $\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*)) > 0$ and therefore

$$\sum_{g \notin H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 \leq 3 \sum_{g \in H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \frac{\epsilon_n}{2},$$

i.e., $\hat{\beta} - \beta^* \in S(3, \frac{\epsilon_n}{2})$. The next proposition provides a lower bound for $\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*))$.

Proposition 4. *On the event $\mathcal{A} \cap \mathcal{B}$ we have*

$$\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*)) \geq c_n \mathbb{E}\left[(\hat{\beta}^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt}))^2\right]$$

with $c_n := \min_{\left\{|x| \leq L\left(\min_{g \in G_n}\{\hat{\beta}_g^{ls}\} M + B\right)\right\} \cap \Theta_{out}} \left\{\frac{\Psi''_{out}(x)}{2}\right\} + \min_{\left\{|x| \leq L\left(\min_{g \in G_n}\{\hat{\beta}_g^{ls}\} M + B\right)\right\} \cap \Theta_{trt}} \left\{\frac{\Psi''_{trt}(x)}{2}\right\}$

Proof. We have

$$\begin{aligned}
&\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*)) \\
&= \mathbb{P}(\Phi_{out}(\hat{\beta}) - \Phi_{out}(\beta^*)) + \mathbb{P}(\Phi_{trt}(\hat{\beta}) - \Phi_{trt}(\beta^*)) \\
&= \mathbb{P}(\Phi_{out}(\hat{\beta}) - \Phi_{out}(\beta^*)) + \mathbb{P}(\Phi_{trt}(\hat{\beta}) - \Phi_{trt}(\beta^*)).
\end{aligned}$$

Recall $\beta = (\alpha, \gamma)^T$ where α are the regression parameters in the outcome model and γ are the regression parameters in the treatment model, and note that

$$\begin{aligned}
&\mathbb{P}(\Phi_{out}(\hat{\beta}) - \Phi_{out}(\beta^*)) \\
&= -\mathbb{E}\left[\mathbb{E}\left(Y|Z_{out}\right)\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)\right] \\
&+\mathbb{E}\left[\psi'_{out}(\alpha^{*T} Z_{out})\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)\right] \\
&+\mathbb{E}\left[\frac{\psi''_{out}(\tilde{\alpha}^T Z_{out})}{2}\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)^2\right],
\end{aligned}$$

where $\tilde{\alpha}^T Z_{out}$ is an intermediate point between $\hat{\alpha}^T Z_{out}$ and $\alpha^{*T} Z_{out}$ given by a second order Taylor expansion of ψ_{out}. Since $\psi'_{out}(\alpha^{*T} Z_{out}) = \mathbb{E}(Y|Z_{out})$ we find

$$\mathbb{P}(\Phi_{out}(\hat{\alpha}) - \Phi_{out}(\alpha^*)) = \mathbb{E}\left[\frac{\psi''_{out}(\tilde{\alpha}^T Z_{out})}{2}\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)^2\right].$$

Besides we have

$$
\begin{aligned}
|\tilde{\alpha}^T Z_{out}| &\leq |\tilde{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}| + |\alpha^{*T} Z_{out}| \\
&\leq \sum_{g=1}^{G_n} |\tilde{\alpha}^{gT} Z_{out}^g - \alpha^{*gT} Z_{out}^g| + \sum_{g=1}^{G_n} |\alpha^{*gT} Z_{out}| \\
&\leq \sum_{g=1}^{G_n} |\hat{\alpha}_n^{gT} Z_{out}^g - \alpha^{*gT} Z_{out}^g| + \sum_{g=1}^{G_n} |\alpha^{*gT} Z_{out}| \\
&\leq \sum_{g=1}^{G_n} \left\|\hat{\alpha}_g - \alpha_g^*\right\|_2 \|Z_{out}^g\|_2 + \sum_{g=1}^{G_n} \left\|\alpha_g^*\right\|_2 \|Z_{out}^g\|_2 ,
\end{aligned}
$$

where the first inequality and second line follows from the triangle inequality, the third line follows because $\tilde{\alpha}^T Z_{out}$ is between $\hat{\alpha}_n^T Z_{out}$ and $\alpha^{*T} Z_{out}$, and the fourth line follows from Hölder's inequality. Applying (H.1), we find

$$
\|Z_{out}\|_2 \leq L\sqrt{d_g} \rightarrow
$$

$$
|\tilde{\alpha}^T Z_{out}| \leq L \left(\sum_{g=1}^{G_n} \left\|\hat{\alpha}_g - \alpha_g^*\right\|_2 \sqrt{d_g} + \sum_{g=1}^{G_n} \left\|\alpha_g^*\right\|_2 \sqrt{d_g} \right).
$$

Then using Lemma 3 and (H.3) we find

$$
|\tilde{\alpha}^T Z_{out}| \leq L \left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} M + B \right) \ a.s.
$$

Moreover, α^* and $\hat{\alpha}$ belong to Λ_{out}, which is a convex set, so we know $\tilde{\alpha} \in \Lambda_{out}$, and therefore, $\tilde{\alpha}^T Z_{out} \in \Theta_{out}$ a.s. It follows that

$$
\mathbb{P}(\Phi_{out}(\hat{\alpha}) - \Phi_{out}(\alpha^*)) \geq c_{1n} \mathbb{E}\left[\left(\hat{\alpha} Z_{out} - \alpha^* Z_{out} \right)^2 \right]
$$

where $c_{1n} := \min_{\left\{ |x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} M + B \right) \right\} \cap \Theta_{out}} \left\{ \frac{\Psi''_{out}(x)}{2} \right\}$.

We can use a similar argument to show

$$
\mathbb{P}(\Phi_{trt}(\hat{\gamma}) - \Phi_{trt}(\gamma^*)) \geq c_{2n} \mathbb{E}\left[\left(\hat{\gamma} Z_{trt} - \gamma^* Z_{trt} \right)^2 \right]
$$

where $c_{2n} := \min_{\left\{|x|\leq L\left(\min_{g\in G_n}\{\hat{\beta}_g^{ls}\}M+B\right)\right\}\cap\Theta_{trt}} \left\{\frac{\Psi''_{trt}(x)}{2}\right\}$.
Therefore,

$$\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{trt}(\beta^*)) \geq (c_{1n}+c_{2n})\mathbb{E}\left[(\hat{\alpha}Z_{out} - \alpha^* Z_{out})^2 + (\hat{\gamma}Z_{trt} - \gamma^* Z_{trt})^2\right]$$

$$\geq (c_{1n}+c_{2n})\mathbb{E}\left[(\hat{\beta}^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt}))^2\right].$$

□

From Proposition 4 and (2.9) we deduce that

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\hat{\beta}_g - \beta_g^*||_2 + c_n\mathbb{E}\left[\left(\hat{\beta}^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt})\right)^2\right]$$

$$\leq 4\lambda_n \sum_{g\in H^*} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\hat{\beta}_g - \beta_g^*||_2 + \frac{\lambda_n}{2}\epsilon_n. \tag{2.10}$$

Let $\Sigma = \mathbb{E}\left[(Z_{out}, Z_{trt})(Z_{out}, Z_{trt})^T\right]$ be the covariance matrix. We have

$$\mathbb{E}\left[\left(\hat{\beta}_n^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt})\right)^2\right] = (\hat{\beta}-\beta^*)^T\Sigma(\hat{\beta}-\beta^*).$$

Because condition $GS(3, \frac{\epsilon_n}{2}, k)$ is satisfied (by assumption) we have

$$c_n(\hat{\beta}-\beta^*)^T\Sigma(\hat{\beta}-\beta^*) \geq c_n k \sum_{g\in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2 - \frac{\epsilon_n}{2}$$

which implies from (2.10) that

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}\left\|\hat{\beta}_g - \beta_g^*\right\|_2 + c_n k \sum_{g\in H^*}\left\|\hat{\beta}_g - \beta_g^*\right\|_2^2$$

$$\leq 4\lambda_n \sum_{g\in H^*} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}\left\|\hat{\beta}_g - \beta_g^*\right\|_2 + \frac{\lambda_n}{2}\epsilon_n + \frac{\epsilon_n}{2}.$$

Then using the Cauchy-Schwarz inequality on the line above we find

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\hat{\beta}_g - \beta_g^*||_2 + c_n k \sum_{g\in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2$$

$$\leq 4\lambda_n \sqrt{\sum_{g\in H^*} \frac{d_g}{(\hat{\beta}_g^{ls})^2}} \sqrt{\sum_{g\in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2} + (\lambda_n + 1)\frac{\epsilon_n}{2}.$$

Now the fact that $2xy \leq tx^2 + y^2/t$ for all $t > 0$ leads to the following inequality (with $x = 2\lambda_n \sqrt{\sum_{g\in H^*} \frac{d_g}{(\hat{\beta}_g^{ls})^2}}$, $y = \sqrt{\sum_{g\in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2}$, and recalling that $\zeta^* = \sum_{g\in H^*} \frac{d_g}{(\hat{\beta}_g^{ls})^2}$):

$$\lambda_n \sum_{g=1}^{G_n} \sqrt{d_g} ||\hat{\beta}_g - \beta_g^*||_2 + c_n k \sum_{g\in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2$$

$$\leq 4t\lambda_n^2 \zeta^* + \frac{1}{t} \sum_{g\in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2 + (\lambda_n + 1)\frac{\epsilon_n}{2}. \tag{2.11}$$

Replacing t by $\frac{1}{c_n k}$ in (2.11) (and dividing by λ_n) we obtain

$$\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2 \leq \frac{4}{c_n k}\lambda_n \zeta^* + \left(1 + \frac{1}{\lambda_n}\right)\frac{\epsilon_n}{2}.$$

What is more, letting $W_g = \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}$, we have

$$\sum_{g=1}^{G_n} W_g \left\| \left(\hat{\beta}_g - \beta_g^*\right) I(\alpha_g^* \neq 0) \right\|_2 \leq \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2.$$

This yields

$$\sum_{g:\alpha_g^*\neq 0} \left\| \left(\hat{\beta}_g - \beta_g^*\right) \right\|_2 \leq \frac{4}{\min_{g:\alpha_g^*\neq 0}\{W_g\} c_n k}\lambda_n \zeta^* + \left(1 + \frac{1}{\lambda_n}\right) \frac{1}{\min_{g:\alpha_g^*\neq 0}\{W_g\} 2n}.$$

Finally we conclude the proof using Proposition 1.

Proof of Proposition 1:

Let $A > \sqrt{2}$. Recall that we have assumed G_n and n are such that $\frac{\log(2G_n)}{n} \leq 1$. We deduce Proposition 1 from the following two lemmas:

Lemma 5. *Let*

$$\lambda_n \geq \left(8\sqrt{2} A L C_{L,B} \sqrt{\frac{\log(2G_n)}{n}}\right) \vee \left(16 A^2 L C_{L,B} \frac{\log(2G_n)}{n}\right)$$

with $A > 1$. Then

$$\mathbb{P}\{\mathcal{A}\} \geq 1 - 2d_{max}(2G_n)^{1-A^2}$$

Lemma 6. *Let*

$$\lambda_n \geq 20AL \left(\underset{(|x| \leq L\kappa_n) \cap \Theta}{max} |\Psi'_{sum}(x)| \right) \sqrt{\frac{2\log(2G_n)}{n}}$$

where $A \geq 1$. Then

$$\mathbb{P}\{\mathcal{B}\} \geq 1 - 2C(2G_n)^{-A^2/2}$$

where we recall $\kappa_n := 17B + \frac{2}{n}$. We can notice that $\mathbb{P}(\mathcal{B})$ tends to 1 as n goes to ∞.

Thus if

$$\lambda_n \geq AKL \left\{ C^*_{L,B} \vee \underset{\{|x| \leq L\kappa_n\} \cap \Theta}{max} |\Psi'_{sum}(x)| \right\} \sqrt{\frac{2\log(2G_n)}{n}}$$

with K chosen such that

$$\lambda_n \geq max(C_1, C_2, C_3)$$

where

$$C_1 := 8\sqrt{2}ALC^*_{L,B}\sqrt{\frac{\log(2G_n)}{n}}$$

$$C_2 := 16A^2LC^*_{L,B}\frac{\log(2G_n)}{n}$$

and

$$C_3 := 20AL \left(\underset{(|x| \leq L\kappa_n) \cap \Theta}{max} |\Psi'_{sum}(x)| \right) \sqrt{\frac{2\log(2G_n)}{n}}$$

then $\mathbb{P}(\mathcal{A} \cap \mathcal{B}) \geq 1 - (2d_{max} + 2C)(2G_n)^{-A^2/2}$. □

Proof of Lemma 3

The proof of Lemma 3 is based on convexity of the loss function and of the penalty, as in Blazère et al. (2014), where the main idea is similar to the one used by Bühlmann and van de Geer (2011) for the lasso to show consistency of the excess risk. Define $t := \frac{M}{M + \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}^{ls}_g|} ||\hat{\beta}_g - \beta^*_g||_2}$ and $\tilde{\beta} := t\hat{\beta} + (1-t)\beta^*$. By convexity of Φ_{sum} and the L_2 norm, in addition to the fact that $\hat{\beta}$ satisfies (2.7), we find

$$\mathbb{P}(\Phi_{sum}(\tilde{\beta}) - \Phi_{sum}(\beta^*)) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}^{ls}_g|} ||\tilde{\beta}_g||_2$$

$$\leq (\mathbb{P}_n - \mathbb{P})(\Phi_{sum}(\beta^*) - \Phi_{sum}(\tilde{\beta})) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\beta_g^*||_2.$$

On the event $\mathcal{A} \cap \mathcal{B}$ we have (from Propositions 2 and 3)

$$\mathbb{P}(\Phi_{sum}(\tilde{\beta}) - \Phi_{sum}(\beta^*)) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\tilde{\beta}_g||_2$$

$$\leq \lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\tilde{\beta}_g - \beta_g^*||_2 + \lambda_n \frac{\epsilon_n}{2} + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\beta_g^*||_2.$$

Because $\mathbb{P}(\Phi_{sum}(\tilde{\beta}) - \Phi_{sum}(\beta^*)) \geq 0$, by adding to both sides of the inequality $2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\beta_g^*||_2$ and by using the triangle inequality, we have

$$\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\tilde{\beta}_g - \beta_g^*||_2 \leq \frac{\epsilon_n}{2} + 4 \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\beta_g^*||_2.$$

Therefore, using (H.3), we have

$$\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\tilde{\beta}_g - \beta_g^*||_2 \leq \frac{\epsilon_n}{2} + 4 \min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} B = \frac{M}{2},$$

i.e.,

$$t \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2 \leq \frac{M}{2},$$

and then the definition of t leads to

$$\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2 \leq M.$$

□

Proof of Lemma 5:

Proof. We have

$$\mathbb{P}(\mathcal{A}^C) \leq$$

$$\sum_{g=1}^{G_n} \mathbb{P} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i} \right) - E \left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt} \right) \right\} \right\|_2^2 > \frac{\lambda_n^2}{4} d_g \right\} \leq$$

$$\sum_{g=1}^{G_n}\sum_{j=1}^{d_g}\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\left\{\left(\frac{Y_i}{sd(Y)}Z^g_{out,i}, A_iZ_{trt,i}\right) - E\left(\frac{Y}{sd(Y)}Z^g_{out}, AZ_{trt}\right)\right\}\right| > \frac{\lambda_n}{2}\right\}. \quad (2.12)$$

We will define random variables $\{W^g_{ij}\}$ with $j = 1, 2$ (more generally, $j = 1, \ldots, d_g$) and $i = 1, \ldots, n$ such that

$$W^g_{i1} := \frac{Y_i}{sd(Y)}Z^g_{out,i} - \mathbb{E}\left(\frac{Y_i}{sd(Y)}Z^g_{out}\right)$$

and

$$W^g_{i2} := A_iZ^g_{trt,i} - \mathbb{E}(A_iZ^g_{out})$$

for $i = 1, \ldots, n$. The random variables $\{W_{ij}\}_{i=1,\ldots,n}$ are independent, identically distributed and centered, and for all $m \geq 2$,

$$\mathbb{E}|W^g_{i1}|^m \leq \sum_{k=0}^{m}\binom{m}{k}\mathbb{E}\left|\frac{Y_i}{sd(Y)}Z_{out,i}\right|^k\left(\mathbb{E}\left|\frac{Y_i}{sd(Y)}Z_{out,i}\right|\right)^{m-k}$$

and

$$\mathbb{E}|W^g_{i2}|^m \leq \sum_{k=0}^{m}\binom{m}{k}\mathbb{E}|A_iZ_{trt,i}|^k(\mathbb{E}|A_iZ_{trt,i}|)^{m-k}.$$

By Jensen's inequality, we obtain

$$\mathbb{E}|W^g_{i1}|^m \leq 2^m \max_{k=1,\ldots,m}\left\{\mathbb{E}\left|\frac{Y_i}{sd(Y)}Z_{out,i}\right|^k\mathbb{E}\left|\frac{Y_i}{sd(Y)}Z_{out,i}\right|^{m-k}\right\}$$

and

$$\mathbb{E}|W^g_{i2}|^m \leq 2^m \max_{k=1,\ldots,m}\{\mathbb{E}|A_iZ_{trt,i}|^k\mathbb{E}|A_iZ_{trt,i}|^{m-k}\}.$$

For all $k \in \mathbb{N}$, by (H.1) and Lemma 4 we have

$$\mathbb{E}\left|\frac{Y_i}{sd(Y)}Z_{out,i}\right|^k \leq L^k k!(C^{out}_{L,B})^k$$

and

$$\mathbb{E}|A_iZ_{trt,i}|^k \leq L^k k!(C^{trt}_{L,B})^k.$$

Therefore $\mathbb{E}|W^g_{ij}|^m \leq m!(2LC^*_{L,B})^m$, where $C^*_{L,B} = \max\{C^{out}_{L,B}, C^{trt}_{L,B}\}$. Hence the conditions are satisfied to apply Bernstein's concentration inequality (Bennett, 1962) with $K = 2LC^*_{L,B}$

and $\sigma^2 = 8(LC^*_{L,B})^2$. Thus we obtain

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} W_{ij}^g\right| > \lambda_n/2\right)$$

$$\leq 2\left(\exp\left(\frac{-n\lambda_n}{16LC^*_{L,B}}\right) + \exp\left(\frac{-n\lambda_n^2}{32(2LC^*_{L,B})^2}\right)\right). \tag{2.13}$$

Finally, from (2.12) and (2.13), we deduce that $\mathbb{P}(\mathcal{A}^c)$ is bounded by

$$2d_{max}G_n\left(\exp\left(\frac{-n\lambda_n}{16LC^*_{L,B}}\right) + \exp\left(\frac{-n\lambda_n^2}{32(2LC^*_{L,B})^2}\right)\right).$$

Therefore if

$$\lambda_n \geq A^2 16LC^*_{L,B}\frac{\log(2G_n)}{n} \vee A8\sqrt{2}LC^*_{L,B}\sqrt{\frac{\log(2G_n)}{n}}$$

with $A > 1$ then

$$\mathbb{P}\{\mathcal{A}^c\} \leq 2d_{max}(2G_n)^{1-A^2}.$$

□

Proof of Lemma 6:

Proof. The proof rests on the following Lemma:

Lemma 7. *Let $R > 0$ be given. Define*

$$Z_R := \sup_{\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq R} \{|(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta))|\}.$$

If $A \geq 1$ then

$$\mathbb{P}\left(Z_R \geq A5DLR\sqrt{\frac{2\log(2G_n)}{n}}\right) \leq 2(2G_n)^{-A^2}$$

where $D := \max\left\{\underset{\{|x|\leq L(R+B)\}\cap\Theta}{max}\{|\Psi'_{out}(x) + \Psi'_{trt}(x)|\}\right\}$.

Proof. Let $R > 0$ be given and β satisfy $\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq R$. Then we know $Z_{R,out} := \sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\alpha_g - \alpha_g^*||_2 \leq R$ (and similarly, $Z_{R,trt} := \sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\gamma_g - \gamma_g^*||_2 \leq R$). Notice that if we change X_i by X_i' while keeping the others fixed then $Z_{out,R}$ is modified by

at most

$$\frac{2}{n}\left(\underset{g\in\{1,\ldots,G_n\}}{min}\left\{|\hat{\beta}_g^{ls}|\right\}/\sqrt{d_g}\right)LR\exp(L(\underset{g\in\{1,\ldots,G_n\}}{min}\left\{|\hat{\beta}_g^{ls}|\right\}R+B)).$$

To see this let

$$\mathbb{P}_n=\frac{1}{n}\sum_{j=1}^{n}1_{X_j,Y_j}$$

and

$$\mathbb{P}'_n=\frac{1}{n}\sum_{j=1}^{n}1_{X_j,Y_j}+1_{X'_j,Y'_j}$$

then we have

$$(\mathbb{P}_n-\mathbb{P})(\Phi_{sum,\Psi}(\beta^*)-\Phi_{sum,\Psi}(\beta))-(\mathbb{P}'_n-\mathbb{P})(\Phi_{sum,\Psi}(\beta^*)-\Phi_{sum,\Psi}(\beta))$$

$$=(\mathbb{P}_n-\mathbb{P})(\Phi_{out,\Psi_{out}}(\alpha^*)-\Phi_{out,\Psi_{out}}(\alpha))-(\mathbb{P}'_n-\mathbb{P})(\Phi_{out,\Psi_{out}}(\alpha^*)-\Phi_{out,\Psi_{out}}(\alpha))$$

$$+(\mathbb{P}_n-\mathbb{P})(\Phi_{out,\Psi_{out}}(\gamma^*)-\Phi_{out,\Psi_{out}}(\gamma))-(\mathbb{P}'_n-\mathbb{P})(\Phi_{out,\Psi_{out}}(\gamma^*)-\Phi_{out,\Psi_{out}}(\gamma))$$

$$=\frac{1}{n}(\Phi_{out,\Psi out}(\alpha^*,Z_{out,i})-\Phi_{out,\Psi_{out}}(\alpha,Z_{out,i})-\Phi_{out,\Psi_{out}}(\alpha^*,Z'_{out,i})+\Phi_{out,\Psi_{out}}(\alpha,Z'_{out,i}))$$

$$+\frac{1}{n}(\Phi_{trt,\Psi trt}(\gamma^*,Z_{trt,i})-\Phi_{trt,\Psi_{trt}}(\gamma,Z_{trt,i})-\Phi_{trt,\Psi_{trt}}(\gamma^*,Z'_{trt,i})+\Phi_{trt,\Psi_{trt}}(\gamma,Z'_{trt,i}))$$

$$\leq\frac{1}{n}|\Psi'(\tilde{\alpha}^TZ_{out,i})||\alpha^{*T}Z_{out,i}-\alpha^TZ_{out,i}|+\frac{1}{n}|\Psi'(\tilde{\alpha}^TZ'_{out,i})||\alpha^{*T}Z'_{out,i}-\alpha^TZ'_{out,i}|$$

$$+\frac{1}{n}|\Psi'(\tilde{\gamma}^TZ_{trt,i})||\gamma^{*T}Z_{trt,i}-\gamma^TZ_{trt,i}|+\frac{1}{n}|\Psi'(\tilde{\gamma}^TZ'_{trt,i})||\gamma^{*T}Z'_{trt,i}-\gamma^TZ'_{trt,i}|$$

where $\tilde{\alpha}Z_{out,i}$ is an intermediate point between $\alpha^TZ_{out,i}$ and $\alpha^{*T}Z_{out,i}$ (using a first order Taylor expansion of the exponential function, as in the proof to Proposition 4). Then, applying (H.1), we find

$$\|Z_{out}\|_2\leq L\sqrt{d_g}\rightarrow$$

$$|\tilde{\alpha}^TZ_{out}|\leq L\left(\sum_{g=1}^{G_n}\left\|\hat{\alpha}_g-\alpha_g^*\right\|_2\sqrt{d_g}+\sum_{g=1}^{G_n}\left\|\alpha_g^*\right\|_2\sqrt{d_g}\right).$$

Then using (H.3) we find

$$|\tilde{\alpha}^TZ_{out}|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right).$$

Similarly, it can be shown that

$$|\tilde{\gamma}^T Z_{trt}| \leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R + B\right).$$

Therefore

$$\begin{aligned}
&(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta)) - (\mathbb{P}'_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta)) \\
&\leq \frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{out}} |\Psi'(\tilde{\alpha}^T Z_{out,i})||\alpha^{*T} Z_{out,i} - \alpha^T Z_{out,i}| \\
&+\frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{out}} |\Psi'(\tilde{\alpha}^T Z'_{out,i})||\alpha^{*T} Z'_{out,i} - \alpha^T Z'_{out,i}| \\
&+\frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z_{trt,i})||\gamma^{*T} Z_{trt,i} - \gamma^T Z_{trt,i}| \\
&+\frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z'_{trt,i})||\gamma^{*T} Z'_{trt,i} - \gamma^T Z'_{trt,i}| \\
&\leq \frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{out}} |\Psi'(\tilde{\alpha}^T Z_{out,i})|\sum_{g=1}^{G_n}||\alpha^* - \alpha||_2||Z_{out}^g||_2 \\
&+\frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{out}} |\Psi'(\tilde{\alpha}^T Z'_{out,i})|\sum_{g=1}^{G_n}||\alpha^* - \alpha||_2||Z_{out}^g||_2 \\
&+\frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z_{trt,i})|\sum_{g=1}^{G_n}||\gamma^* - \gamma||_2||Z_{trt}^g||_2 \\
&+\frac{1}{n}\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z'_{trt,i})|\sum_{g=1}^{G_n}||\gamma^* - \gamma||_2||Z_{trt}^g||_2 \\
&\leq \frac{2}{n}\left(\min_{g\in\{1,\ldots,G_n\}}\left\{|\hat{\beta}_g^{ls}|\right\}/\sqrt{d_g}\right)LR\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{out}}\{|\Psi'_{out}(x)|\} \\
&+\frac{2}{n}\left(\min_{g\in\{1,\ldots,G_n\}}\left\{|\hat{\beta}_g^{ls}|\right\}/\sqrt{d_g}\right)LR\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{trt}}\{|\Psi'_{trt}(x)|\}
\end{aligned}$$

$$= \frac{4}{n} M_w LRD$$

where $M_w = \left(\underset{g \in \{1,\dots,G_n\}}{min} \left\{ |\hat{\beta}_g^{ls}| \right\} / \sqrt{d_g} \right)$.
We can apply McDiarmid's inequality (also called the bounded difference inequality) to Z_R and obtain

$$\mathbb{P}(Z_R - \mathbb{E}Z_R \geq u) \leq \exp\left(-\frac{nu^2}{8M_w^2 R^2 L^2 D^2} \right).$$

Therefore if $\lambda_n \geq ADM_w LR\sqrt{\frac{8 \log 2G_n}{n}}$ with $A > 0$ then

$$\mathbb{P}(Z_{R,out} - \mathbb{E}Z_{R,out} \geq \lambda_n) \leq (2G_n)^{-A^2}. \tag{2.14}$$

Now we have to bound the mean $\mathbb{E}Z_R$. To do this, we need the Symmetrization theorem and the contraction principle (see Appendix A of Blazère et al. (2014)), and then let $\epsilon_1, \dots, \epsilon_n$ be a Rademacher sequence independent of $Z_{out,1}, \dots, Z_{out,n}$ and $Z_{trt,1}, \dots Z_{trt,n}$ and let $S_R := \{\beta \in \mathbb{R}^p : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\beta_g - \beta_g^*||_2 \leq R\}$. Then by the Symmetrization theorem and the Contraction principle (since ψ is D-lipschitz on the compact set S_R) we have

$$\mathbb{E}Z_R \leq 4D\mathbb{E}\left(\sup_{\beta \in S_R} \frac{1}{n} \sum_{i=1}^{n} |\epsilon_i (\beta^{*T} Z_i - \beta^T Z_i)| \right)$$

$$\leq 4DR\mathbb{E}\left(\max_{g \in \{1,\dots,G_n\}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{|\hat{\beta}_g^{ls}| ||Z_i^g||_2}{\sqrt{d_g}} \right| \right),$$

where the last bound follows from Holder's inequality. By applying the theorem below from Blazère et al. (2014) that's a consequence of Hoeffding's inequality, we obtain

$$\mathbb{E}\left(\max_{g \in \{1,\dots,G_n\}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{|\hat{\beta}_g^{ls}| ||Z_i^g||_2}{\sqrt{d_g}} \right| \right) \leq M_w L \sqrt{\frac{2 \log(2G_n)}{n}}.$$

Theorem. (Blazère et al., 2014) *Let $X_1, \dots, X_n$ be independent random variables on χ and $f_1, \dots, f_n$ real-valued functions on χ which satisfies for all $j = 1, \dots, p$ and all $i = 1, \dots, p$ and all $i = 1, \dots, n$*

$$Ef_j(X_i) = 0, |f_j(X_i)| \leq a_{ij}.$$

Then

$$E\left(\underset{1 \leq j \leq p}{max} \left| \sum_{i=1}^{n} f_j(X_i) \right| \right) \leq \sqrt{2 \log(2p)} \underset{1 \leq j \leq p}{max} \sqrt{\sum_{i=1}^{n} a_{ij}^2}.$$

It follows that

$$\mathbb{E}Z_R \leq 4M_w RLD\sqrt{\frac{2\log(2G_n)}{n}}. \tag{2.15}$$

□

Thus from (2.14) and (2.15) we know that if $A \geq 1$ then

$$\mathbb{P}\left(Z_R \geq ADM_w LR\left(\sqrt{\frac{8\log 2G_n}{n}} + \sqrt{\frac{2\log(2G_n)}{n}}\right)\right) \leq (2G_n)^{-A^2}$$

for all $R > 0$. □

Split up

$$\left\{\beta \in \mathbb{R}^p : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq M\right\},$$

where $M = 8\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}\right)B + \epsilon_n$, into two sets which are

$$E_1 = \left\{\beta : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq \epsilon_n\right\}$$

and

$$E_2 = \left\{\beta : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq M\right\}$$

$$\subseteq \bigcup_{j=1}^{j_n}\left\{\beta : 2^{j-1}\epsilon_n < \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq 2^j\epsilon_n\right\}$$

where $j_n := [\log_2(nM)] + 1$ is the smaller integer such that $2j_n\epsilon_n \geq M$. We recall that

$$v_n := \frac{(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta))}{\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 + \epsilon_n}$$

and to simplify notation let

$$\alpha(\beta, \beta^*) := (\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta))$$

and

$$\Omega(t) := \max\{ \underset{\{|x|\leq L\left(\min_{g\in G_n}\{\hat{\beta}_g^{ls}\}R+B\right)\}\cap\Theta_{out}}{max} \{|\Psi'_{out}(x)|\}, \underset{\{|x|\leq L\left(\min_{g\in G_n}\{\hat{\beta}_g^{ls}\}R+B\right)\}\cap\Theta_{trt}}{max} \{|\Psi'_{trt}(x)|\}\}.$$

Let $A \geq 1$. Recall that $\kappa_n := 17B + \frac{2}{n} = 2M + B$. On the event E_1,

$$\mathbb{P}\left(\underset{\beta\in E_1}{sup} |v_n(\beta,\beta^*)| \geq A10L\Omega(L\kappa_n)\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq \mathbb{P}\left(\underset{\beta\in E_1}{sup} |\alpha(\beta,\beta^*)| \geq A10L\Omega(L\kappa_n)\epsilon_n\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq \mathbb{P}\left(\underset{\beta\in E_1}{sup} |\alpha(\beta,\beta^*)| \geq A5L\Omega(L(\epsilon_n+B))\epsilon_n\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

given that $2M \geq \epsilon_n$. From Lemma 7 with $R = \epsilon_n$ we deduce

$$\mathbb{P}\left(\underset{\beta\in E_1}{sup} |v_n(\beta,\beta^*)| \geq A10L\Omega(L\kappa_n)\epsilon_n\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq 2(2G_n)^{-A^2}. \tag{2.16}$$

On the event E_2, using the same type of argument as (2.16) with $R = 2^j\epsilon_n$ (given that $2M \geq 2^j\epsilon_n$) for all $j = 1,\ldots,j_n$, we find

$$\mathbb{P}\left(\underset{\beta\in E_2}{sup} |v_n(\beta,\beta^*)| \geq A10L\Omega(L\kappa_n)\epsilon_n\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq j_n 2(2G_n)^{-A^2}.$$

Finally we have

$$\leq C'2(2G_n)^{-\frac{A^2}{2}} \tag{2.17}$$

where C' is a constant (because $j_n = [\log_2(nM)]+1$ and $n << G_n$) and the result of Lemma 6 follows from (2.16) and (2.17) with $C = 1 + C'$. □

2.8.3 Supplement C: Additional Simulations

In Chapter 2.5, simulations are presented comparing the performance of GLiDeR, the two-stage model averaged double robust estimator proposed by Cefalu et al. (2017) (which we abbreviate as "MADR"), two standard doubly robust estimators – one using all covariates ("saturated method") and another which selects covariates via "backward selection" (p-stay = 0.05) on the outcome model – and a non-doubly robust method using the adaptive lasso on only the outcome model to select covariates and estimate the average causal effect for ten scenarios ("Scenarios 1–10" presented in Chapter 2.5.1) with $p = 10$ covariates (independent and correlated) and sample size $n = 500$. Additional simulation scenarios are presented here exploring the effects of adjusting the number of covariates and sample size in Scenarios 1–9.

Ratio of mean squared error (MSEs) of the doubly robust average causal treatment effect of GLiDeR, backward selection, MADR, and adaptive lasso (denominator) relative to the saturated variable selection method (numerator) for 5 independent covariates and sample size $n = 500$ and for 10 independent covariates and sample size $n = 250$ are shown in Tables 2.4 and 2.5, respectively.

Table 2.4: Ratio of MSE (saturated model MSE / alternative method MSE) for each scenario with independent data, 5 covariates, and sample size n = 500 over 1,000 Monte Carlo datasets.

Scenario	GLiDeR MSE Ratio	Backward Selection MSE Ratio	MADR MSE Ratio	Adaptive Lasso MSE Ratio
1	**1.09**	1.01	**1.10**	**1.11**
2	**1.08**	1.00	**1.10**	**1.11**
3	**2.77**	1.00	**2.88**	**3.00**
4	**1.00**	1.00	**1.01**	**0.66**
5	**1.63**	**1.04**	**1.62**	**1.64**
6	**16.82**	**0.91**	**18.15**	**16.00**
7	**1.10**	**1.03**	**1.09**	**1.13**
8	**1.21**	**1.03**	**1.21**	**1.29**
9	**1.02**	**1.02**	**1.02**	**1.02**

Bold indicates significant difference (5% significance level) between MSEs (testing equality) from the saturated method (full model) vs. the alternative method using the paired t-test.

The same results are shown for 25 independent covariates and sample size $n = 500$ in Table 2.6, but results were not calculated for MADR due to the relatively large number of covariates. The MSE ratios with 5 covariates (Table 2.4) are slightly smaller for all

Table 2.5: Ratio of MSE (saturated model MSE / alternative method MSE) for each scenario with independent data, 10 covariates, and sample size n = 250 over 1,000 Monte Carlo datasets.

Scenario	GLiDeR MSE Ratio	Backward Selection MSE Ratio	MADR MSE Ratio	Adaptive Lasso MSE Ratio
1	**1.13**	**1.03**	**1.15**	**1.18**
2	**1.12**	**1.04**	**1.14**	**1.14**
3	**3.59**	**1.17**	**3.77**	**3.97**
4	**1.04**	**1.02**	**1.05**	**0.90**
5	**1.81**	**0.93**	**1.79**	**1.78**
6	**24.31**	**1.27**	**24.77**	**21.81**
7	**1.19**	**1.06**	**1.16**	**1.21**
8	**1.39**	**1.10**	**1.38**	**1.46**
9	**1.11**	**1.08**	**1.12**	**1.12**

Bold indicates significant difference (5% significance level) between MSEs (testing equality) from the saturated method (full model) vs. the alternative method using the paired t-test.

Table 2.6: Ratio of MSE (saturated model MSE / alternative method MSE) for each scenario with independent data, 25 covariates, and sample size n = 500 over 1,000 Monte Carlo datasets.

Scenario	GLiDeR MSE Ratio	Backward Selection MSE Ratio	Adaptive Lasso MSE Ratio
1	**1.13**	**1.12**	**1.15**
2	**1.18**	**1.17**	**1.20**
3	**3.57**	**3.49**	**4.06**
4	**1.06**	**1.06**	**0.98**
5	**1.94**	**1.80**	**1.89**
6	**24.71**	**9.41**	**22.30**
7	**1.23**	**1.11**	**1.25**
8	**1.45**	**1.34**	**1.53**
9	**1.13**	**1.11**	**1.15**

Bold indicates significant difference (5% significance level) between MSEs (testing equality) from the saturated method (full model) vs. the alternative method using the paired t-test.

methods and scenarios compared to 10 covariates (Table 2.5 in main manuscript) as the alternative methods (GLiDeR, backward selection, MADR, and adaptive lasso) are generally more efficient than the saturated method when there are more irrelevant variables. This is further seen for GLiDeR, adaptive lasso, and backward selection with 25 covariates (Table 2.6) as these methods obtain much greater MSE ratios for all scenarios than with 5 and 10 covariates. When the sample size is cut in half ($n = 250$) with 10 covariates (Table 2.5), the

MSE ratios increase in nearly all scenarios for all alternative methods. In other words, the MSE ratios are further away from 1 for all methods and scenarios when the sample size is halved, which seems to suggest the gap in performance between methods is increased with a smaller sample size. As in the main manuscript, the adaptive lasso only considers the outcome model and under-selects an important confounder weakly related to the outcome but strongly associated to the treatment in Scenario 4 and is less efficient than the saturated method even with 25 covariates.

Results are presented below testing generalized cross-validation (GCV) and k-fold cross-validation (kCV) for $k = 2, 5$, and 10 folds on the outcome model for Scenarios 1–10 with 10 independent covariates for Scenarios 1–9 and 100 independent covariates for Scenario 10 and a sample size of 500 for all scenarios. GCV is performed as discussed in Chapter 2.3.4 and k-fold cross-validation chooses the tuning parameter value λ^* as the value λ yielding the smallest average mean squared prediction error across the k test folds. Performance is generally similar for all procedures, but GCV demonstrates the best performance overall at estimating the causal treatment effect in these scenarios, and also has a computational advantage over kCV (especially for larger k) as it requires the method to be computed only once on the data. Consequently, we recommend using GCV over kCV for model selection.

Table 2.7: Comparison of tuning parameter selection procedures.

	GCV			2-fold			5-fold			10-fold		
Scenario	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD
1	0.0081	0.00	0.09	0.0081	0.00	0.09	0.0081	0.00	0.09	0.0081	0.00	0.09
2	0.0090	0.00	0.09	0.0089	0.00	0.09	0.0090	0.00	0.10	0.0090	0.00	0.09
3	0.0085	0.00	0.09	0.0091	0.00	0.10	0.0094	0.00	0.10	0.0091	0.00	0.10
4	0.0100	0.00	0.10	0.0100	0.00	0.10	0.0100	0.00	0.10	0.0101	0.01	0.10
5	0.4008	-0.04	0.63	0.4368	-0.05	0.66	0.4371	-0.06	0.66	0.4433	-0.06	0.66
6	0.0958	0.00	0.31	0.1350	-0.01	0.37	0.1057	0.00	0.33	0.1275	0.00	0.36
7	0.7040	-0.05	0.84	0.7175	-0.05	0.85	0.7213	-0.05	0.85	0.7268	-0.05	0.85
8	0.0143	0.01	0.12	0.0141	0.01	0.12	0.0144	0.01	0.12	0.0143	0.01	0.12
9	0.0117	0.00	0.11	0.0117	0.00	0.11	0.0118	0.00	0.11	0.0118	0.00	0.11
10	0.0603	0.00	0.26	0.0638	-0.07	0.24	0.0644	-0.08	0.24	0.0906	-0.15	0.26

Table 2.8: Covariates selected (average across 1000 samples) by GLiDeR. Though $p = 100$ covariates are considered for Scenario 10, only results for the first two irrelevant variables (X_9 and X_{10}) are shown here.

Scenario	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0.07	0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04
2	1.00	0.12	1.00	1.00	0.04	0.04	0.04	0.04	0.03	0.04
3	1.00	1.00	0.07	0.06	0.06	0.02	0.01	0.01	0.01	0.01
4	1.00	1.00	1.00	1.00	1.00	0.05	0.04	0.04	0.04	0.04
5	0.28	0.28	0.80	0.79	0.80	0.06	0.05	0.05	0.06	0.06
6	0.16	0.17	0.08	0.09	0.16	0.02	0.02	0.02	0.02	0.03
7	0.51	0.11	1.00	0.92	0.09	0.06	0.06	0.08	0.06	0.06
8	1.00	1.00	0.59	0.06	0.06	0.05	0.04	0.05	0.05	0.06
9	1.00	0.06	1.00	1.00	0.04	0.05	0.04	0.06	0.05	0.06
10	1.00	0.99	1.00	1.00	0.21	0.20	0.21	0.19	0.01	0.01

Table 2.9: Bootstrap 95% percentile confidence interval coverage rates by GLiDeR for all scenarios with sample size $n = 500$ and $p = 10$ covariates (except Scenario 10, which has $p = 100$ covariates) across $1,000$ Bootstrap samples. Note that correlated covariates are not considered for Scenario 10.

Scenario	Independent Covariates Coverage Rate	Correlated Covariates Coverage Rate
1	95.1	95.2
2	94.4	94.8
3	94.3	94.1
4	93.5	95.2
5	95.2	94.5
6	94.3	95.5
7	94.4	91.9
8	95.0	93.6
9	94.8	94.7
10	97.1	-

Table 2.10: Covariates (potential confounders) considered in the lung transplant registry. Each variable is continuous or binary. The mean and standard deviation (if continuous) or frequency and proportion (if binary) of each covariate for BLT and SLT is also shown.

Name	Description	BLT Mean (sd)/ N (%)	SLT Mean (sd)/ N (%)
	Patient characteristics		
AgeP	Age (yrs)	63.6 (2.9)	64.2 (3.1)
BmiP	Body Mass Index	24.5 (7.4)	24.8 (7.4)
DiabP	Diabetes	64 (13%)	41 (9%)
HgtP	Height (cm)	169.8 (9.1)	169.3 (9.3)
O2amt	Oxygen delivered	4.07 (3.07)	3.43 (1.93)
Karn	Karnofsky score > 60	155 (31%)	188 (42%)
LAS	Lung allocation score	35.8 (7.6)	34.0 (3.6)
WhtP	Race (white)	455 (92%)	416 (94%)
SexP	Gender (female)	211 (43%)	208 (47%)
LifeS	Life support ventilator needed	27 (5%)	4 (1%)
Vent	Assisted ventilation needed	68 (14%)	49 (11%)
Vol	Center volume	94.5 (66.5)	71.3 (45.8)
Walk	6 minute walking distance	746.7 (390.7)	719.2 (322.2)
O2rest	Oxygen needed at rest	31 (6%)	36 (8%)
	Donor characteristics		
AgeD	Age (yrs)	36.3 (14.4)	33.7 (14.4)
BlckD	Race (black)	92 (19%)	87 (20%)
BmiD	Body Mass Index	26.0 (5.2)	25.4 (4.9)
Cig	History of cigarette use	74 (15%)	57 (13%)
CMV	Positive cytomegalovirus (CMV) test	302 (61%)	266 (60%)
Cod	Cause of death - traumatic brain injury	224 (45%)	243 (55%)
DiabD	Diabetes	38 (8%)	24 (5%)
ExpD	Expanded donor	65 (13%)	52 (12%)
HgtD	Height (cm)	175.5 (9.4)	175.3 (9.2)
SexD	Gender (female)	146 (30%)	135 (30%)
Dist	Donor to treatment center distance	206 (243.8)	203.3 (246.9)
Po2	Lung PO2	387.2 (148.4)	364.5 (151.3)
	Other characteristics		
Allo	Local or regional (vs. national) allocation	146 (30%)	114 (26%)
HgtR	Height ratio	1.03 (0.05)	1.04 (0.05)
Isch	Ischemic time	5.5 (1.6)	4.0 (1.4)
SexM	Matching gender	125 (25%)	131 (30%)
RaceM	Matching race	330 (67%)	274 (62%)

Chapter 3

Variable selection and estimation in causal inference using Bayesian spike and slab priors

3.1 Introduction

Inferring the causal effect of a treatment, exposure, or intervention (hereafter referred to as "treatment") on some outcome or response is often the primary goal of a study. Randomizing treatment assignment is the gold standard for estimating causal treatment effects but is unethical, infeasible, or not cost-effective in many situations. When treatment is not randomized, confounding variables – those associated with both treatment and outcome – can induce bias in the treatment effect estimator if ignored. There are many ways to adjust for confounding variables; many approaches involve modeling the outcome as a function of treatment and covariates or modeling the treatment as a function of covariates (or both).

Assuming all measured covariates contain all confounding variables, an unbiased estimator of the causal treatment effect can be obtained from a model that correctly specifies the conditional mean of the outcome as a function of treatment and covariates. Controlling for all measured covariates prevents confounder omission and, therefore, protects against bias. However, adjusting for covariates that are unrelated to the outcome can increase the variance of the treatment effect estimator without reducing bias, so the "all-inclusive" approach

can be suboptimal for estimating treatment effects, especially when many measured covariates are under consideration. Another approach is to use variable selection and adjust for only the covariates that are related to the outcome (which includes confounding variables). Traditional variable selection techniques, such as the lasso, select covariates based on their associations in only the outcome model and may not select important confounding variables that are weakly related to outcome but strongly associated with treatment. Using variable selection on only the outcome (or treatment) model can, therefore, bias the treatment effect estimator.

Bayesian model averaging (BMA) proposes taking a weighted average of the effect estimates across models with different covariates included where the weights are determined by the posterior model probability (Raftery, Madigan & Hoeting, 1997). However, like traditional variable selection, standard BMA tends to prioritize models which include covariates strongly associated with outcome and may assign significant weight to models that include only a subset of the necessary confounders, resulting in biased treatment effect estimation (Crainiceanu, Dominici and Parmigiani, 2008). Crainiceanu, Dominici and Parmigiani (2008) introduce a two-stage BMA method that forces strong predictors of treatment that are identified in a first stage to be included in the outcome model in a second stage, and then strong predictors of outcome that are not identified in stage one are identified in stage two. Wang et al. (2012) propose Bayesian Adjustment for Confounding (BAC), a Bayesian model averaging method on the outcome model with an informative prior obtained separately from the treatment model (see Zigler and Dominici (2014) for related Bayesian methods that select variables for propensity score estimation). BAC contains a prior dependence parameter, ω, ranging from 1 to ∞ that links the treatment model to the outcome model. If $\omega = \infty$, all covariates with associations in the treatment model are forced into the outcome model, whereas $\omega = 1$ treats the two models independently and is equivalent to standard BMA on the outcome model.

Two approaches have been predominantly used to select ω: (1) setting ω equal to ∞ as the default and (2) selecting ω data-adaptively to minimize mean squared error (MSE) or other criterion. Each approach is problematic.

Setting $\omega = \infty$ targets the set of covariates associated with treatment or outcome. However, we observe that BAC with $\omega = \infty$ can have high inclusion probabilities for irrelevant covariates that are unrelated to outcome and treatment, particularly in smaller

sample sizes. This can lead to inefficient estimators in moderate sample sizes compared to other variable selection approaches which target the same set of covariates (i.e., all variables associated with treatment or outcome). Furthermore, for a given sample size, selecting all covariates related to treatment and outcome may not be the set which leads to the most efficient estimator of the average causal effect. In Chapter 3.3.4, the variance and bias of the treatment effect estimator is derived, and we find that adjusting for covariates that are strongly related to treatment but unrelated (or very weakly related) to outcome can increase the asymptotic variance of the treatment effect estimator without substantially reducing its asymptotic bias. Consequently, to minimize MSE of the treatment effect estimator, it may be necessary to use models which do not include covariates related only to treatment (or those related to treatment but are very weakly related to the outcome) when estimating the treatment effect.

However, developing a data-adaptive approach to select ω to minimize MSE has proved challenging. Lefebvre, Atherton and Talbot (2014) proposed using cross validation or the bootstrap to choose ω with the aim of minimizing MSE of the treatment effect estimator, but they found that the performance of these procedures was sensitive to the underlying data generating mechanism and suggested that alternative approaches should be investigated. Even if cross validation or the bootstrap could be reliably used to choose ω, such methods can be computationally intensive with large datasets. Further, BAC requires calculating the Bayesian Information Criterion at each posterior draw, which cannot be calculated when the number of potential confounders in the model exceeds the sample size.

In this chapter, we propose the Spike and Slab Causal Estimator (SSCE) and Bilevel SSCE (BSSCE), novel Bayesian methods that simultaneously consider models for outcome and treatment and use spike and slab priors on the covariate coefficients to encourage variable selection based on associations in both the outcome and treatment models. SSCE aims to minimize treatment effect bias by controlling only for covariates that are related to outcome or treatment (and removing irrelevant ones), while BSSCE adjusts for the subset of the covariates which minimize MSE of the treatment effect estimator. The proposed methods, which are adapted from the formulation of the Bayesian group lasso with spike and slab priors Xu and Ghosh (2015), are implemented using fast Gibbs samplers that perform well with a large number of covariates, even when the number of covariates is greater than the sample size.

3.2 Preliminaries

3.2.1 Estimation of causal treatment effects

Suppose an observational study yields outcomes $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ with corresponding binary treatment indicators $\mathbf{A} = \{A_1, \ldots, A_n\}$ for independent subjects $1, \ldots, n$, and we are interested in estimating the average causal effect of a binary treatment A on outcome Y, defined as

$$\Delta = E\{Y(1) - Y(0)\},$$

where $Y(1)$ and $Y(0)$ denote potential outcomes that would be observed from an arbitrary individual in the population if given treatment ($A = 1$) and control ($A = 0$), respectively. Other measures of the causal treatment effect, such as the average effect of treatment on the treated, could also be applied here. If we are willing to assume consistency and strong ignorability, i.e., that measured covariates $\mathbf{X} = X_1, \ldots, X_p$ include all confounding variables so that $\{Y(0), Y(1)\} \perp A|\mathbf{X}$, then

$$E\{E(Y|A = 1, \mathbf{X})\} = E[E\{Y(1)|A = 1, \mathbf{X}\}] = E[E\{Y(1)|\mathbf{X}\}] = E\{Y(1)\},$$

and similarly, $E\{E(Y|A = 0, \mathbf{X})\} = E\{Y(0)\}$. This implies $\Delta = E\{E(Y|A = 1, \mathbf{X}) - E(Y|A = 0, \mathbf{X})\}$, so a correctly specified model for $E(Y|A, \mathbf{X})$ can be used to consistently estimate Δ. If we assume a linear regression model for $E(Y|A, \mathbf{X})$,

$$Y_i|\mathbf{X}_i, A_i, \beta, \sigma^2 \sim N\{\mu(\mathbf{X}_i, A_i, \beta), \sigma^2\}, i = 1, \ldots, n, \tag{3.1}$$

where

$$\mu(\mathbf{X}_i, A_i; \beta) = \beta_0 + \beta_A A_i + \beta^T \mathbf{X}_i,$$

a natural estimator for Δ is

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \{\mu(\mathbf{X}_i, A_i = 1; \hat{\beta}) - \mu(\mathbf{X}_i, A_i = 0; \hat{\beta})\} = \hat{\beta}_A.$$

It is common to adjust for all measured covariates $\mathbf{X}$ when estimating β_A. Assuming $\mathbf{X}$ contains all confounding variables, adjusting for all measured covariates ensures adjustment for all confounders, which protects against bias in the treatment effect estimator that could

otherwise emerge if a confounder is omitted. But since adjusting for measured covariates that are unrelated to $E(Y|A, \mathbf{X})$ can increase the variability of the treatment effect estimator without reducing its bias, some type of dimension reduction to select covariates associated with $E(Y|A, \mathbf{X})$ is necessary to achieve unbiased, efficient estimation of β_A.

3.2.2 Bayesian spike and slab lasso (BSSL)

The lasso Tibshirani (1996) is a popular variable selection technique that can perform well when p is large, possibly even greater than n. Though the standard Bayesian lasso shrinks covariate coefficients toward zero, it does not yield posterior estimates that are exactly zero. Xu and Ghosh (2015) use spike and slab priors on the model coefficients to propose the *Bayesian Spike and Slab Lasso (BSSL)*, a hierarchical formulation of the Bayesian lasso that allows draws from the posterior of each covariate coefficient to be exactly zero. Particularly, the prior on β is

$$\beta_j|\sigma^2, \tau_j^2 \overset{ind.}{\sim} (1-\pi_0)N(0, \sigma^2\tau_j^2) + \pi_0\delta_0(\beta_j), j = 1, \ldots, p,$$

where $\delta_0()$ is a point mass at zero and π_0 is the prior probability that a covariate coefficient is zero; conjugate priors can be used for σ^2 (Inverse Gamma), π_0 (Beta), and τ_j^2 (Gamma). These priors control the amount of shrinkage of the covariate coefficients. Throughout, we assume covariates $\mathbf{X}$ are standardized to have marginal mean zero and unit variance so that covariate selection is invariant to the scale of the covariates.

BSSL can be used to select variables for $\mu(\mathbf{X}, A; \beta)$ and simultaneously estimate β_A, but it ignores the relationship between covariates and treatment. There are no problems using BSSL asymptotically, but in finite samples, BSSL may not select important confounding variables that are weakly related to the outcome, even if they are strongly associated with the treatment. The next section proposes a novel framework that aims to reduce bias in finite samples by encouraging selection of covariates that are associated with treatment assignment.

3.3 Spike and slab causal estimation methodology

3.3.1 Simultaneous modeling of outcome and treatment

We begin by specifying a probit model for the conditional probability that subject i receives treatment ($A_i = 1$),

$$P(A_i = 1|\mathbf{X}_i) = \Phi(\gamma_0 + \gamma^T \mathbf{X}_i),$$

where $\Phi()$ is the cumulative distribution function of a standard normal random variable. An equivalent formulation of the probit model states that there exists A_i^* such that $A_i = 1$ if $A_i^* > 0$ and $A_i = 0$ if $A_i^* \leq 0$, where A_i^* is an unobserved (latent) variable that is normally distributed with mean $\gamma_0 + \gamma^T \mathbf{X}_i$ and unit variance. Such an assumption allows us to model A^* in the way that Y is modeled in BSSL, but without the term for the treatment effect, β_A.

To select covariates based on their associations with both outcome and treatment assignment, we define a new vector $\mathbf{O} = (\mathbf{Y}, \mathbf{A}^*)^T$ by stacking the outcomes and latent treatment variables, and the corresponding design matrix

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X}_{out} & 0 \\ 0 & \mathbf{X}_{trt} \end{pmatrix},$$

consisting of the design matrices $\mathbf{X}_{out} = \{\mathbf{1}_n, \mathbf{A}, \mathbf{X}_1, \ldots, \mathbf{X}_p\}$ for the outcome model and $\mathbf{X}_{trt} = \{\mathbf{1}_n, \mathbf{X}_1, \ldots, \mathbf{X}_p\}$ for the treatment model. We note that $\mathbf{Z}$ contains two columns that are associated with each covariate, a crucial fact for the joint selection technique we introduce in the next section. With this notation, the likelihood can be written as

$$\mathbf{O}|\mathbf{Z}, \alpha, \sigma^2 \sim N_{2n}(\mathbf{Z}\alpha, \Sigma_O), \tag{3.2}$$

with $\alpha = (\beta_0, \beta_A, \beta_1, \ldots, \beta_p, \gamma_0, \gamma_1, \ldots, \gamma_p)^T$ and Σ_O a $2n \times 2n$ diagonal matrix with σ^2 as the first n elements of the diagonal and 1 as the last n elements of the diagonal.

The variable selection techniques which we propose use a modified Bayesian group lasso approach. The group lasso is an extension of the lasso to allow selection of predetermined groups of variables (Yuan and Lin, 2006). For example, if X_i and X_j were related in such a way that we would want to include or exclude both variables simultaneously in the model, we could use the group lasso and group together the regression coefficients that correspond

to X_i and X_j (β_i and β_j); then after grouping the other $p-2$ coefficients in the model so that all p coefficients belong to exactly one group, the group lasso will force all coefficients in a group will be either all zero or all nonzero, meaning β_i and β_j will simultaneously be zero or nonzero. If all groups are of size one, the group lasso and lasso estimators are equivalent.

For our problem, we have two models (one for outcome and one for treatment assignment) and want each covariate to be included or excluded simultaneously from both models (i.e., want β_j and γ_j to be both zero or both nonzero). That is, we want the coefficients in the model for $\mathbf{O}|\mathbf{Z}, \alpha, \sigma^2$ corresponding to the covariates $(\mathbf{X}_j^T, \mathbf{0})$ and $(\mathbf{0}, \mathbf{X}_j^T)$ to be included or excluded simultaneously. We, therefore, use the idea of the group lasso to form p groups of size 2, with each group k containing the outcome and treatment model coefficients corresponding to covariate k (as in Chapter 2):

$$\text{Group } 1 = \{\beta_1, \gamma_1\}, \text{Group } 2 = \{\beta_2, \gamma_2\} \ldots, \text{Group } p = \{\beta_p, \gamma_p\}.$$

3.3.2 Spike and slab causal estimator (SSCE)

Using a similar idea to the Bayesian spike and slab group lasso proposed by Xu and Ghosh (2015) (an extension of BSSL), we propose the following prior with the likelihood in (3.2):

$$\left(\begin{array}{c} \beta_j \\ \gamma_j \end{array} \right) \Bigg| \sigma^2, \tau_j^2 \sim (1-\pi_0) N \left\{ \left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \tau_j^2 \sigma^2 & 0 \\ 0 & \tau_j^2 \end{array} \right) \right\} + \pi_0 \delta_0 \left\{ \left(\begin{array}{c} \beta_j \\ \gamma_j \end{array} \right) \right\}, j = 1, \ldots, p.$$

As in Xu and Ghosh (2015), conjugate priors are used for σ^2 (Inverse Gamma) and π_0 (Beta), the prior probability that $(\beta_j, \gamma_j)^T = (0,0)^T$. Flat priors are used for β_0, β_A, and γ_0 since we do not want to shrink these coefficients toward zero. We also recommend setting τ_j^2 to a large enough positive value so that the Normal slab on the covariate coefficients is effectively non-informative (we used $\tau_j^2 = 1000$ for all j in our simulations; see Chapter 3.3.3 for more details). To implement this model, which we call the Spike and Slab Causal Estimator (SSCE), a fast Gibbs sampler is used to generate samples from the posterior distribution of the treatment effect β_A and other parameters (after standardizing covariates); all priors are conjugate, so full conditionals are easily derived and implemented. The full conditionals used in the Gibbs samplers are given in Chapter 2.8.1, and R code for implementation of SSCE (and BSSCE, which is proposed in Chapter 3.3.4) is available at

https://github.com/drbkoch/SSCE.

3.3.3 A motivating example for SSCE and choice of τ_j^2

In Chapter 3.7.2, a simplified example is presented to illustrate the properties and behavior of SSCE. Figure 3.1 plots the posterior probability that β_j is zero for this example as a function of its least squares estimate ($\hat{\beta}_j^{ls}$) using BSSL (represented by the black line) and SSCE; the posterior probability β_j is zero using SSCE depends on the magnitude of the maximum likelihood estimate of γ_j under the full model (denoted $\hat{\gamma}_j$), and each colored line represents a different value of this estimate using SSCE.

As $\hat{\gamma}_j$ is increased, the posterior probability that β_j is zero decreases and eventually reaches zero, thereby increasing the inclusion probability of confounders in SSCE. Additionally, when covariate j is irrelevant (i.e., $\hat{\beta}_j^{ls}$ and $\hat{\gamma}_j$ are near zero), the posterior probability β_j is zero is larger with SSCE, meaning irrelevant covariates will be selected less frequently with SSCE compared to BSSL.

The posterior distribution of β_j for fixed τ_j^2 is also derived for the simplified example presented in Chapter 3.7.2. There is a finite-sample bias for β_j, with magnitude equal to $\frac{\hat{\beta}_j^{ls}}{1+n\tau_j^2}$, that may increase bias in our estimator for the treatment effect, $\hat{\beta}_A$. To remedy such a problem, we set each τ_j^2 to a common large value that makes the bias negligible in the estimates of each covariate coefficient and minimized shrinkage to zero.

3.3.4 Bilevel spike and slab causal estimator (BSSCE)

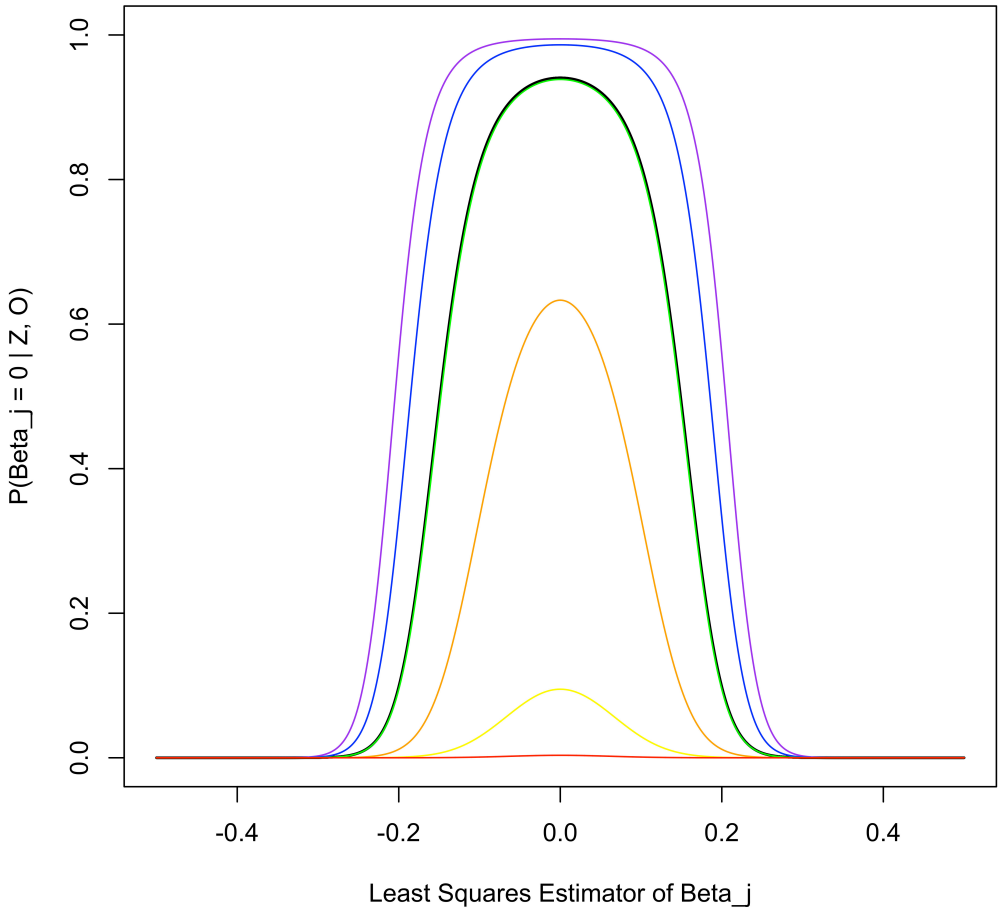
Adjusting for covariates that are strongly related to treatment but unrelated to outcome may increase the variance of the treatment effect estimator without reducing its bias. The asymptotic variance of the least squares treatment effect estimator (i.e., $\hat{\beta}_A$) with covariates $\mathbf{X}_{-j}$, where $\mathbf{X}_{-j}$ is the vector of covariates excluding the jth covariate, is

$$\mathrm{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}) = \frac{1}{n} \frac{\sigma^2_{Y|A,\mathbf{X}_{-j}}}{\sigma^2_{A|\mathbf{X}_{-j}}},$$

and with covariates $\mathbf{X}_{-j}$ and covariate X_j, the asymptotic variance is

$$\mathrm{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}, X_j) = \frac{1}{n} \frac{\sigma^2_{Y|A,\mathbf{X}_{-j}} - \sigma^2_{X_j|\mathbf{X}_{-j},A}\, \beta_j^2}{\sigma^2_{A|\mathbf{X}_{-j}} - \sigma^2_{X_j|\mathbf{X}_{-j}}\, \eta_j^2},$$

Figure 3.1: $P(\beta_j = 0|\mathbf{O}, \mathbf{Z})$ as a function of the least squares estimate of β_j under orthogonal outcome and treatment design matrices. The colors denote the posterior probability β_j is zero using the proposed method for different values of $\hat{\gamma}_j$; purple, blue, green, orange, yellow, and red (i.e., from top to bottom of figure at least squares estimate of β_j equal to 0) respectively represent $\hat{\gamma}_j$ equal to 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30. The black line denotes the posterior probability β_j is zero under BSSL. For this Figure, $n = 250$, $\pi_0 = 0.5$, $\sigma^2 = 1$, and $\tau_j^2 = 1$.



where $\sigma^2_{L|\mathbf{R}}$ denotes the residual variance when fitting a linear model on L with covariates $\mathbf{R}$, and β_j is the true coefficient corresponding to covariate X_j in the outcome and η_j is the coefficient corresponding to X_j from regressing A on $\mathbf{X}$ using a linear model. The asymptotic bias of $\hat{\beta}_A$ with covariates $\mathbf{X}_{-j}$ (and X_j excluded) is

$$\text{Bias}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}) = \beta_j \eta_j \frac{\sigma^2_{X_j|\mathbf{X}_{-j}}}{\sigma^2_{A|\mathbf{X}_{-j}}}.$$

Note that because we assume that we have measured all confounders, the bias with covariates $\mathbf{X}_{-j}$ and X_j is 0. Using a first-order Taylor series expansion and assuming that X_j is normally distributed, η_j can be approximated $\frac{\gamma_j}{\sqrt{2\pi}}$. We use this approximation throughout. Adjusting for covariate X_j that is unrelated to outcome (i.e., with small $|\beta_j|$) but strongly associated with treatment (i.e., with large $|\gamma_j|$) will have a small change in bias (no change if $\beta_j = 0$) but potentially large increase in variance.

To reduce MSE (equal to the variance plus the square of the bias) of the treatment effect estimator, we propose BSSCE, which has the following spike and slab prior on each group of coefficients:

$$\begin{pmatrix} \beta_j \\ \gamma_j \end{pmatrix} \Bigg| \sigma^2, \tau_j^2 \sim (1 - \pi_0) D + \pi_0 \delta_0 \left\{ \begin{pmatrix} \beta_j \\ \gamma_j \end{pmatrix} \right\},$$

where D is a two-dimensional distribution with density function $f(\beta_j, \gamma_j)$ that is equal to the density of the normally distributed slab in SSCE except for values (β_j, γ_j) such that

$$\text{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}) - \text{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}, X_j) + \text{Bias}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j})^2 < 0,$$

where the density is set to zero. That is, values of β_j, γ_j for which MSE would be lower if β_j and γ_j were set to zero (covariate j were removed from the model) are not included in the support. Instead, we include point mass at $\beta_j = 0$ and $\gamma_j = 0$ to ensure that $f(\beta_j, \gamma_j)$ is a valid density. We call this a bi-level model because it is reminiscent of the bi-level group lasso proposed by Xu and Ghosh (2015), in which selection is first done between and then within groups. At the first level, the spike in the prior (δ_0) removes covariates irrelevant to both treatment and outcome as in SSCE. However, in BSSCE the second "level" of selection is still at the group level, but specifically reduces MSE. This "level" of selection

occurs within the part of the prior denoted by D here. The result of the restricted support is to shrink any β_j, γ_j on covariates whose inclusion in the model increase the MSE (due to weak association with outcome) to zero, as indicated by the second point mass described above. Coefficients that are equal to zero using SSCE are also zero using BSSCE. However, coefficients that are non-zero for SSCE may be zero for BSSCE. Computationally, BSSCE uses the same Gibbs sampler as SSCE, except BSSCE implements a second stage that sets non-zero coefficient values from SSCE which increase MSE to zero.

An empirical Bayes approach is used to estimate $\sigma^2_{Y|A,\mathbf{X}}$, $\sigma^2_{A|\mathbf{X}}$, $\sigma^2_{X_j|A,\mathbf{X}_{-j}}$, and $\sigma^2_{X_j|\mathbf{X}_{-j}}$. To estimate $\sigma^2_{Y|A,\mathbf{X}_{-j}}$, the standard lasso estimator, call it $\hat{\beta}^{lasso}$, is obtained for the coefficients in the regression of $\mathbf{X}$ and A on Y. Our estimate of $\sigma^2_{Y|A,\mathbf{X}_{-j}}$ for all covariates j such that $\hat{\beta}^{lasso}_j = 0$ is then

$$\hat{\sigma}^2_{Y|A,\mathbf{X}_{-j}} = \frac{1}{n - \sum_k I(\hat{\beta}^{lasso}_k \neq 0)} \sum_{i=1}^{n} (Y_i - \hat{\beta}^{lasso}_0 - \hat{\beta}^{lasso}_1 X_{1i} - \cdots - \hat{\beta}^{lasso}_p X_{pi} - \hat{\beta}^{lasso}_A A_i)^2.$$

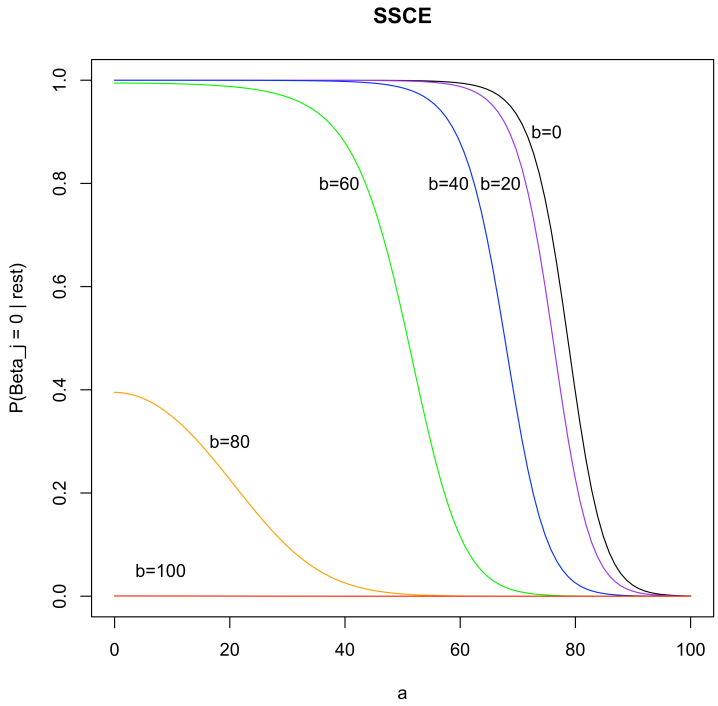
Next, the lasso estimator for the coefficients in the regression of $\mathbf{X}_{-j}$ and A on Y is obtained (call it $\hat{\beta}^{lasso,-j}$) for all covariates j such that $\hat{\beta}^{lasso}_j \neq 0$; the estimate of $\sigma^2_{Y|A,\mathbf{X}_{-j}}$ is then

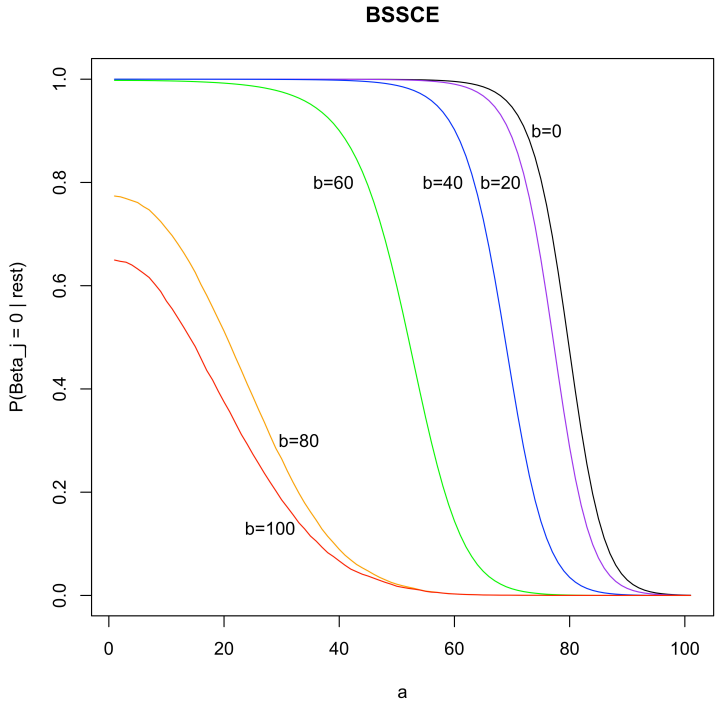
$$\hat{\sigma}^2_{Y|A,\mathbf{X}_{-j}} = \frac{1}{n - \sum_k I(\hat{\beta}^{lasso,-j}_k \neq 0)} \sum_{i=1}^{n} (Y_i - \hat{\beta}^{lasso,-j}_0 - \hat{\beta}^{lasso,-j}_1 X_{1i} - \cdots - \hat{\beta}^{lasso,-j}_p X_{pi} - \hat{\beta}^{lasso,-j}_A A_i)^2.$$

Such an estimation technique is computationally inexpensive and allows for estimation when p is greater than n.

The probit model used for the treatment assignment assumes there exists a latent variable (which determines the treatment assignment indicator) with a linear mean and unit error variance. This latent variable (called A^* in Chapter 3.3.1) can be viewed as the treatment assignment on a continuous scale, in which case $\sigma^2_{A|\mathbf{X}}$ (the error variance assuming treatment assignment follows a linear regression model as a function of $\mathbf{X}$) would approximately equal one. We, therefore, use $\hat{\sigma}^2_{A|\mathbf{X}_{-j}} = 1$ for all j. Similarly, assuming covariates are independent and not caused by the treatment, then since covariates are standardized to have unit variance, we use $\hat{\sigma}^2_{X_j|A,\mathbf{X}_{-j}} = 1$ and $\hat{\sigma}^2_{X_j|\mathbf{X}_{-j}} = 1$. These parameter estimates performed well in simulations, even when covariates were not independent.

Figure 3.2: $P(\beta_j = 0|\text{rest})$ as a function of a and b with large slabs (i.e., τ_j^2 large), where a and b are proportional to the correlation between the jth covariate and the residual vectors without the jth covariate in the outcome and treatment models, respectively. For this figure, $n = 250$, $\pi_0 = 0.5$, and $\sigma^2 = 1$.





3.3.5 Covariate inclusion probability for SSCE and BSSCE

In Chapter 3.7.3, the conditional probability that β_j (and γ_j) is zero under SSCE is derived assuming only that covariates are standardized (as in Chapter 3.2.2). Figure 3.2 plots this probability as a function of a and b, where a and b are proportional to the correlation between the jth covariate and the residual vectors without the jth covariate in the outcome and treatment models, respectively. The probability β_j is zero decreases as a function of $\frac{a^2}{\sigma^2} + b^2$, meaning β_j is less likely to be zero if covariate j has stronger associations with the residuals in the outcome or treatment model. Thus, SSCE is more likely than BSSL to include confounders with weak associations to outcome but strong associations to treatment, meaning it should have less bias than BSSL. Additionally, irrelevant covariates that have small $\frac{a^2}{\sigma^2}+b^2$ will have small probability of having non-zero coefficients, meaning the variability in the estimator for β_A should be reduced using SSCE compared to a model that adjusts for such irrelevant covariates. However, covariates that are strongly related to treatment will always be included when using SSCE, even if they are unrelated to outcome. BSSCE, on the other hand, allows covariates only related to treatment to have inclusion probability less than one, which can be seen in Figure 3.2, where covariate inclusion probabilities are derived for BSSCE as functions of a and b using Monte Carlo integration.

3.4 Simulations

Table 3.1: Covariates $X_1, \ldots, X_p$ are generated with mean μ_x and variance V_x, where $\text{Cor}(X_i, X_j) = \rho$ for $i \neq j, i, j \leq 20$. and $\text{Cor}(X_i, X_j) = 0$ for $i \neq j, i, j > 20$, and treatment indicators and corresponding outcomes are generated from Bernoulli(expit(μ_A)) and Normal(μ_Y, V_Y), respectively, where expit(x) = $\frac{exp(x)}{1+exp(x)}$, for the following scenarios.

Scenario	μ_x	V_x	ρ	μ_A	μ_Y	V_Y
1	1	4	0	$0.2X_1 - 2X_2 + X_5 - X_6 + X_7 - X_8$	$2X_1 + 0.2X_2 + 5X_3 + 5X_4$	4
2	0	1	0.5	$-2X_1 - 2X_2 - 2X_3 + X_4 + X_5 - 0.2X_6 + 2X_7 + X_8 - 0.5X_9 + X_{10}$	$\sum_{j=1}^{9} 0.2X_j + 0.5X_{10} - 5X_{11} - 5X_{12}$	1
3	0	1	0.5	$-2X_1 - 2X_2 - 2X_3 + X_4 + X_5 - 0.2X_6 + 2X_7 + X_8 - 0.5X_9 + X_{10} - 2\sum_{j=13}^{16}(-1)^j X_j$	$\sum_{j=1}^{9} 0.2X_j + 0.5X_{10} - 5X_{11} - 5X_{12}$	1

To evaluate the performance of SSCE and BSSCE, we considered three different data-generating scenarios described in Table 4.1. In Scenario 1, covariates are independent, and there is a single confounder that is weakly associated with the outcome and strongly related to the treatment (X_2). In Scenarios 2 and 3, the covariates are correlated, and there are multiple confounders weakly associated with outcome. There are no covariates related to

treatment but unrelated to outcome in Scenario 2, while in Scenario 3, some covariates $(X_{13}, \ldots, X_{16})$ are related to treatment but not outcome. We vary the sample size in all scenarios and vary the number of irrelevant covariates in Scenario 1.

We compare SSCE and BSSCE to BSSL – with $\tau_j^2 = 1000$ to remove shrinkage bias – and to BAC with $\omega = \infty$. We used Inverse Gamma(0.1, 0.1) and Beta(1, 1) distributions as priors for σ^2 and π_0, respectively, in the simulation. Standard `R` programming was used to implement SSCE, BSSCE, and BSSL, while the `R` package `bacr` was used to implement BAC. Note that treatment indicators are generated according to a logistic regression model as assumed in the analysis by BAC but not by the other methods, which instead posit a probit model for treatment assignment.

Simulations are replicated over 500 Monte Carlo (MC) samples, and 5,000 Markov Chain MC samples were used per chain for each method, where the MC standard error for these chain lengths was estimated (using the `R` package `mcmcse`) to range from approximately 0.005-0.01. To compare variable selection performance, the probability of inclusion for each covariate is calculated as the proportion of posterior draws for which its outcome regression coefficient is non-zero. The MC bias, standard error, and MSE of the treatment effect estimator are calculated for each method using the posterior mean of β_A. 95% credible intervals for the treatment effect are estimated for each method using the 2.5% and 97.5% quantiles of the posterior distribution of β_A. Code to reproduce the simulation is available at https://github.com/drbkoch/SSCE.

Table 4.2 shows the average inclusion probabilities for the first ten covariates in Scenario 1 for different sample sizes (n) and total number of covariates (p). Using the Bayesian lasso on the outcome model with spike and large slab priors (BSSL with fixed large τ_j^2), the average inclusion probability for the confounder weakly related to outcome (X_2) is near zero for all considered combinations of n and p, except with the largest ratio of n to p ($n = 500, p = 50$), where the inclusion probability for X_2 is 0.37. Using SSCE, however, yields an average inclusion probability close or equal to one for the weak confounder for all considered n and p. Average inclusion probabilities of the weak confounder are similar using BAC compared to SSCE. However, the average inclusion probabilities of irrelevant covariates are approximately zero using SSCE, even with twice as many covariates as subjects ($n = 250, p = 500$), and much smaller compared to those of BAC.

By including irrelevant covariates less frequently, SSCE decreases the variability and

Table 3.2: Covariate inclusion probabilities in Scenario 1 for the first 10 covariates under numerous combinations of n (sample size) and p (number of covariates). X_1 is a confounder strongly associated with outcome and weakly associated with treatment; X_2 is a confounder weakly associated with outcome, $X_5 - X_8$ are only associated with treatment, and X_9 and X_{10} are irrelevant.

		Weak	Only Related			Only Related				
	Confounder	Confounder	to Outcome			to Treatment			Irrelevant	
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
$n = 100, p = 50$										
BSSL	1	0.02	1	1	0	0	0	0	0	0
BAC	1	0.95	1	1	0.83	0.80	0.83	0.82	0.34	0.34
SSCE	1	0.88	1	1	0.09	0.08	0.09	0.13	0	0
BSSCE	1	0.41	1	1	0	0	0	0.01	0	0
$n = 150, p = 50$										
BSSL	1	0.04	1	1	0	0	0	0	0	0
BAC	1	1	1	1	0.89	0.89	0.90	0.91	0.40	0.41
SSCE	1	1	1	1	0.40	0.37	0.39	0.43	0	0
BSSCE	1	0.71	1	1	0.02	0.01	0.01	0.02	0	0
$n = 200, p = 50$										
BSSL	1	0.06	1	1	0	0	0	0	0	0
BAC	1	1	1	1	0.96	0.97	0.93	0.96	0.45	0.46
SSCE	1	1	1	1	0.77	0.76	0.77	0.78	0	0
BSSCE	1	0.85	1	1	0.03	0.03	0.03	0.03	0	0
$n = 250, p = 50$										
BSSL	1	0.08	1	1	0	0	0	0	0	0
BAC	1	1	1	1	1	1	0.99	1	0.51	0.50
SSCE	1	1	1	1	0.94	0.94	0.94	0.94	0	0
BSSCE	1	0.91	1	1	0.06	0.05	0.06	0.05	0	0
$n = 500, p = 50$										
BSSL	1	0.37	1	1	0	0	0	0	0	0
BAC	1	1	1	1	1	1	1	1	0.10	0.12
SSCE	1	1	1	1	1	1	1	1	0	0
BSSCE	1	0.98	1	1	0.25	0.25	0.25	0.28	0	0
$n = 250, p = 100$										
BSSL	1	0.06	1	1	0	0	0	0	0	0
BAC	1	1	1	1	1	1	1	1	0.10	0.12
SSCE	1	1	1	1	0.77	0.76	0.77	0.77	0	0
BSSCE	1	0.91	1	1	0.05	0.04	0.05	0.04	0	0
$n = 250, p = 250$										
BSSL	1	0.04	1	1	0	0	0	0	0	0
BAC	*	*	*	*	*	*	*	*	*	*
SSCE	1	1	1	1	0.66	0.66	0.67	0.67	0	0
BSSCE	1	0.91	1	1	0.04	0.03	0.04	0.03	0	0
$n = 250, p = 500$										
BSSL	1	0.02	1	1	0	0	0	0	0	0
BAC	*	*	*	*	*	*	*	*	*	*
SSCE	1	1	1	1	0.59	0.56	0.57	0.57	0	0
BSSCE	1	0.91	1	1	0.03	0.02	0.03	0.02	0	0

*Too many covariates for BAC

MSE of the treatment effect estimator compared to BAC for all considered combinations of n and p in Scenario 1 except the largest considered ratio of n to p ($n = 500$ and $p = 50$), where the variability and MSE of the treatment effect estimator are similar between BAC and SSCE (see Table 4.3). BSSCE, which aims to reduce MSE of the treatment effect estimator, leads to much smaller inclusion probabilities for covariates related only to treatment in Scenario 1 compared to SSCE and BAC, and nearly identical inclusion probabilities for irrelevant covariates as SSCE. Though the weak confounder is selected less often with BSSCE than with SSCE and BAC in Scenario 1, BSSCE still achieves similar 95% CI coverage of the treatment effect as SSCE and BAC, and yields a significantly smaller MSE of the treatment effect than all considered methods for all considered ratios of n to p (except $n = 100$ and $p = 50$ - the smallest ratio - where BSSCE and SSCE perform similarly).

Table 3.3: MC Bias, standard error (SE), MSE, and 95% credible interval (CI) coverage probability for the treatment effect estimators.

	Bias	SE	MSE	95% CI Coverage
Scenario 1				
$n = 100, p = 50$				
BSSL	-0.419	0.438	0.367	0.826
BAC	-0.059	0.741	0.551	0.904
SSCE	-0.025	0.554	0.307	0.934
BSSCE	-0.160	0.532	0.308	0.918
$n = 150, p = 50$				
BSSL	-0.403	0.373	0.301	0.782
BAC	0.005	0.566	0.320	0.938
SSCE	0.014	0.475	0.226	0.952
BSSCE	-0.041	0.434	0.189	0.938
$n = 200, p = 50$				
BSSL	-0.393	0.323	0.258	0.722
BAC	0.019	0.461	0.213	0.942
SSCE	-0.000	0.424	0.179	0.968
BSSCE	-0.014	0.360	0.130	0.950
$n = 250, p = 50$				
BSSL	-0.373	0.308	0.234	0.694
BAC	0.023	0.421	0.177	0.928
SSCE	0.009	0.395	0.156	0.958
BSSCE	0.001	0.328	0.107	0.954
$n = 500, p = 50$				
BSSL	-0.232	0.291	0.138	0.754
BAC	0.022	0.292	0.085	0.956
SSCE	0.014	0.291	0.085	0.946
BSSCE	-0.002	0.240	0.058	0.962
$n = 250, p = 100$				
BSSL	-0.389	0.297	0.239	0.674
BAC	0.041	0.431	0.187	0.936
SSCE	0.009	0.386	0.149	0.944
BSSCE	0.000	0.327	0.107	0.950
$n = 250, p = 250$				
BSSL	-0.403	0.285	0.244	0.648
BAC	*	*	*	*
SSCE	0.007	0.378	0.142	0.946
BSSCE	-0.001	0.325	0.105	0.950
$n = 250, p = 500$				
BSSL	-0.410	0.281	0.247	0.646
BAC	*	*	*	*
SSCE	-0.000	0.369	0.136	0.948
BSSCE	0.002	0.318	0.101	0.944
Scenario 2				
$n = 100, p = 30$				
BSSL	-0.233	0.455	0.260	0.916
BAC	0.002	0.657	0.431	0.934
SSCE	-0.141	0.477	0.247	0.950
BSSCE	-0.179	0.463	0.246	0.942
$n = 250, p = 30$				
BSSL	-0.248	0.308	0.156	0.864
BAC	-0.026	0.363	0.132	0.938
SSCE	0.009	0.353	0.125	0.954
BSSCE	-0.100	0.311	0.107	0.946
$n = 500, p = 30$				
BSSL	-0.256	0.258	0.132	0.754
BAC	-0.012	0.256	0.066	0.950
SSCE	-0.011	0.261	0.068	0.956
BSSCE	-0.001	0.260	0.067	0.958
Scenario 3				
$n = 100, p = 30$				
BSSL	-0.196	0.438	0.230	0.912
BAC	-0.024	0.666	0.443	0.944
SSCE	-0.158	0.439	0.217	0.930
BSSCE	-0.173	0.436	0.220	0.932
$n = 250, p = 30$				
BSSL	-0.175	0.296	0.118	0.888
BAC	0.035	0.399	0.160	0.944
SSCE	-0.044	0.358	0.130	0.928
BSSCE	-0.079	0.300	0.096	0.930
$n = 500, p = 30$				
BSSL	-0.153	0.216	0.070	0.894
BAC	0.011	0.268	0.072	0.940
SSCE	0.039	0.273	0.076	0.942
BSSCE	-0.022	0.212	0.046	0.954

*Too many covariates for BAC

In Scenario 2, where covariates are correlated and there are numerous weak confounders, the BSSL estimator is biased, as expected, and yields the largest MSE of all methods for scenarios where $n = 250$ and $n = 500$. The bias using SSCE is much larger than that

of BAC with $n = 100$. In this situation, the inclusion probabilities for the confounding variables are larger using BAC compared to SSCE (and larger for SSCE compared to BSSL). However, BAC displays much larger inclusion probabilities for irrelevant variables and much larger variability in the treatment effect estimator compared to the other methods, and consequently has the largest MSE of all methods when $n = 100$ and $p = 30$. Even with bias, SSCE still achieves treatment effect credible interval coverage probability at the nominal level, even larger than BAC. When the sample size is increased to $n = 250$, the MSE using SSCE is still smaller than that of BAC, and BSSCE obtains the smallest MSE of all methods. With $n = 500$, the MSEs using SSCE, BSSCE, and BAC are similar, which is expected since there are no covariates strongly related to treatment but unrelated (or very weakly related) to outcome. That is, SSCE, BSSCE, and BAC all target the same set of covariates in Scenario 2, so we would expect that with moderate sample size their performance would be similar.

In Scenario 3, which is similar to Scenario 2 except that there are covariates directly related to treatment but not directly related to outcome, SSCE reduces the MSE of the treatment effect estimator over BSSL and BAC for all sample sizes considered, except with $n = 500$ where SSCE and BAC achieve a similar MSE. BSSCE yields a similar MSE to SSCE with $n = 100$, and a significantly smaller MSE than the other methods with $n = 250$ and $n = 500$ because BSSCE includes those covariates which are only related to treatment (which lead to increased variance of the treatment effect estimator but no difference in bias) with much lower probability.

3.5 Application

In critical care resolving hypotensive episodes (HEs) in a timely manner is crucial to minimizing end organ damage (Lee et al., 2012). The procedures for treating HEs vary, and there is evidence that certain treatments could be associated with a shorter HE duration for patients in critical care (Lee et al., 2012). Data on patients treated in intensive care units (ICUs) were obtained from the publicly available Multi-parameter Intelligent Monitoring in Intensive Care III (MIMIC-III) database to infer the average causal effect of fluid resuscitation compared to vasoactive therapy on HE duration. MIMIC-III contains descriptive de-identified clinical data (demographics, vital signs, laboratory tests, medications, etc.)

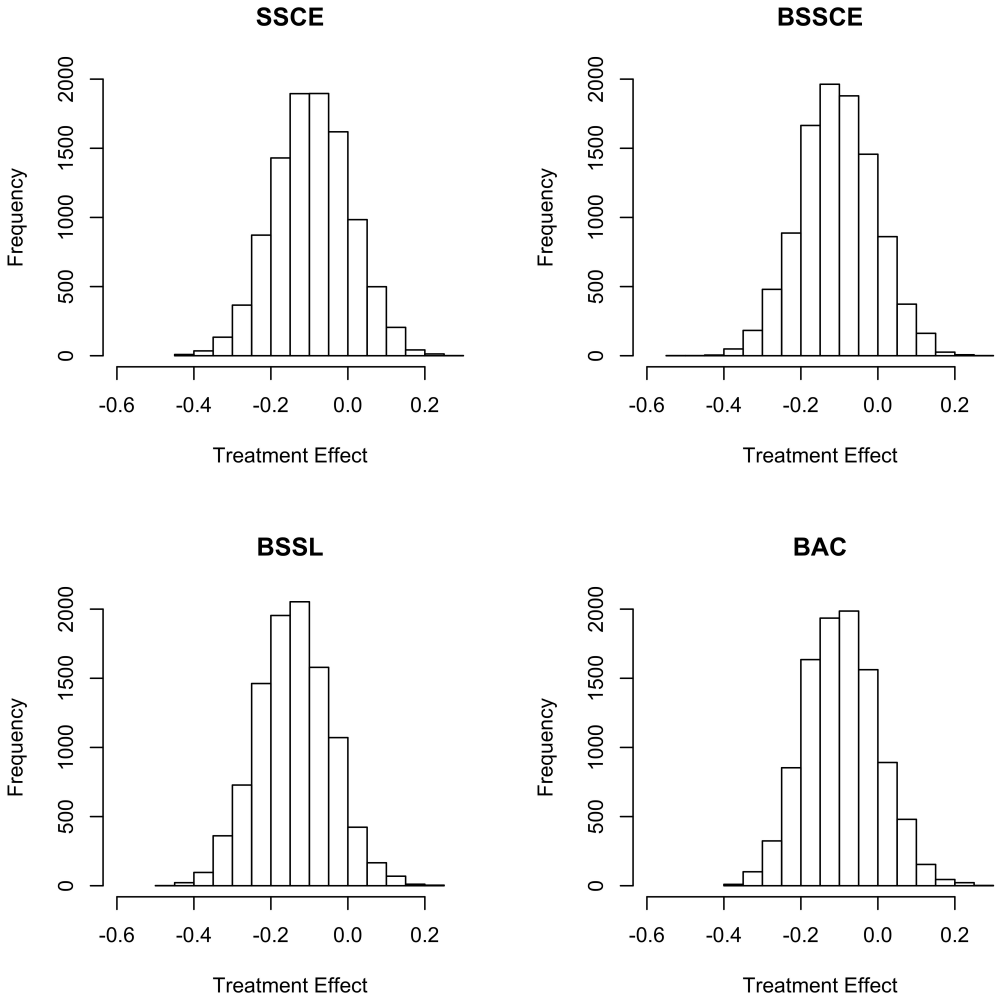
from approximately 50,000 adult patients admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012 (Johnson et al., 2016).

Patients from the MetaVision data management system were included in the analyses if they experienced a HE, which was defined based on mean arterial pressure (MAP) measurements generally recorded every 10-15 minutes, and received vasoactive therapy or fluid resuscitation. Following Lee et al. (2012) the beginning of a HE was defined as the time of two consecutive MAP measurements $\leq$ 60mm Hg, preceded by two consecutive MAP measurements $>$ 60mm Hg. The end of a HE was then defined as the first time that two consecutive MAP measurements $>$ 60mm Hg. Only data on the first HE for each subject was collected so that observations are independent. Vasoactive therapy was defined as an initiation or a dosage increase of dobutamine, dopamine, epinephrine, norepinephrine, phenylephrine or vasopressin during the HE. Fluid resuscitation was defined as at least one infusion of either any volume of colloids or a bolus of isotonic crystalloid.

Table 3.4: Covariates considered in the application. For continuous covariates, the mean and standard deviation (SD) by treatment status is provided, and for categorical covariates, the number observed (N) and percentage (%) by treatment status is given. Inclusion probabilities for SSCE, BSSCE, BSSL, and BAC are also shown.

	Vasoactive therapy Mean/N (SD/%)	Fluid resuscitation Mean/N (SD/%)	Inclusion probabilities			
			SSCE	BSSCE	BSSL	BAC
Mean MAP 3hrs prior to treatment	68.0 (9.6)	66.1 (6.8)	1	1	0.999	1
SAPSII	72.1 (14.1)	64.7 (14.5)	1	0.662	0.001	1
MAP at treatment	56.4 (13.3)	58.1 (11.5)	0	0	0.013	0.885
Liver disease	66 (9.4%)	39 (20.0%)	0	0	0	0.653
Surgical ICU	209 (29.7%)	68 (34.9%)	0	0	0	0.520
Cardiac Surgery Recovery Unit	92 (13.1%)	33 (16.9%)	0	0	0	0.356
Creatinine value	1.6 (1.5)	1.4 (1.3)	0	0	0.001	0.297
Alcohol abuse	53 (7.5%)	24 (12.3%)	0	0	0	0.190
Hypothyroidism	90 (12.8%)	28 (14.4%)	0	0	0	0.160
Medical ICU	79 (11.2%)	7 (3.6%)	0	0	0	0.145
Rheumatoid arthritis	24 (3.4%)	9 (4.6%)	0	0	0	0.142
Age	66.7 (13.7)	65.5 (16.0)	0	0	0	0.125
Sex	428 (60.9%)	107 (54.9%)	0	0	0	0.108
Blood anemias	5 (0.7%)	5 (2.6%)	0	0	0	0.103
Peripheral vascular disorders	75 (10.7%)	18 (9.2%)	0	0	0	0.088
Congestive heart failure	210 (29.9%)	46 (23.6%)	0	0	0	0.085
Renal failure	33 (4.7%)	7 (3.6%)	0	0	0.001	0.084
Paralysis	11 (1.6%)	4 (2.1%)	0	0	0	0.083
Fluid and electrolyte disorders	341 (48.5%)	80 (41.0%)	0	0	0	0.079
Chromic pulmonary disease	167 (23.8%)	36 (18.5%)	0	0	0	0.070
Coronary Care Unit	93 (13.2%)	28 (14.4%)	0	0	0	0.063
Weight loss	46 (6.5%)	16 (8.2%)	0	0	0	0.060
AIDS	3 (0.4%)	1 (0.5%)	0	0	0	0.060
Valvular disease	158 (22.5%)	42 (21.5%)	0	0	0	0.053
Other neurological disorders	74 (10.5%)	14 (7.2%)	0	0	0	0.050
Urine output 3hrs prior to treatment	39.6 (693.8)	17.5 (184.0)	0	0	0	0.050
Diabetes, uncomplicated	182 (25.9%)	48 (24.6%)	0	0	0	0.050
Cardiac arrhythmias	284 (40.4%)	74 (37.9%)	0	0	0	0.048
Drug abuse	25 (3.6%)	10 (5.1%)	0	0	0	0.047
Solid tumor without metastasis	101 (14.4%)	22 (11.3%)	0	0	0	0.047
Lymphoma	17 (2.4%)	4 (2.1%)	0	0	0	0.043
Diabetes, complicated	49 (7.0%)	11 (5.6%)	0	0	0	0.040
Psychoses	29 (4.1%)	9 (4.6%)	0	0	0	0.038
Coagulopathy	162 (23.0%)	49 (25.1%)	0	0	0	0.037
Hypertension, uncomplicated	327 (46.5%)	91 (46.7%)	0	0	0	0.028
Peptic ulcer disease excluding bleeding	12 (1.7%)	6 (3.1%)	0	0	0	0.024
Deficiency anemias	137 (19.5%)	35 (17.9%)	0	0	0	0.020
Pulmonary circulation disorders	71 (10.1%)	24 (12.3%)	0	0	0	0.015

Figure 3.3: Estimated posterior distributions of β_A (causal effect of fluid resuscitation vs. vasoactive therapy on HE duration in minutes, log-transformed) for each method.



The potential confounders are described in Table 4.4. After removing cases with missing covariate values, $n = 898$ observations remained in the analyses with $p = 38$ potential confounders. The outcome of interest was time from treatment to the end of the HE (in minutes and log-transformed). Because these data involve human subjects, the data use agreement does not permit the authors to directly release or archive the data but it may be requested from https://mimic.physionet.org/; code to reproduce these analyses is available from https://github.com/drbkoch/SSCE

Figure 3.3 shows the estimated posterior distribution of β_A ($A = 1$ if vasoactive therapy is used and $A = 0$ if fluid resuscitation is used) using SSCE, BSSCE, BSSL with fixed large τ_j^2, and BAC. All methods yield credible intervals (using the 0.025 and 0.975 quantiles of

the posterior distribution of β_A) that contain zero, indicating no significant difference in HE duration for patients receiving vasoactive therapy compared to fluid resuscitation. Using SSCE, the estimated posterior mean of β_A is -0.09 (95% credible interval = (-0.27, 0.09)), while the estimated posterior mean of β_A using BSSCE is -0.11 (95% credible interval = (-0.30, 0.08)). The estimated posterior mean of β_A using BSSL is -0.14 (95% credible interval = (-0.32, 0.04)), while BAC estimates the posterior mean of β_A to be -0.10 (95% credible interval = (-0.27, 0.09)).

We can see from Table 4.4 that many more covariates have non-zero inclusion probability when using BAC compared to the spike and slab methods, which is consistent with our simulation results. The SAPSII score, which estimates severity of disease, has its outcome coefficient equal to zero in nearly all posterior samples when using BSSL, but has a non-zero coefficient in all of the samples when using SSCE, while all other posterior inclusion probabilities are similar between the two methods (see Table 4.4). The differences in adjustment for the SAPSII score most likely explain why BSSL suggests a more favorable treatment effect than SSCE. When using BSSCE, the SAPSII score has a non-zero coefficient in 66.2% of posterior samples. The difference in inclusion probabilities between SSCE and BSSCE suggests the SAPSII score is either only related to treatment, or a confounder strongly related to treatment and weakly related to outcome that could decrease MSE when setting its coefficient to zero according to the formula in Chapter 3.3.4.

3.6 Discussion

We have proposed two novel Bayesian methods for variable selection and estimation in causal inference that simultaneously model the outcome and treatment assignment using spike and slab priors on the model coefficients. By simultaneously modeling the outcome and treatment assignment, the proposed methods can identify confounding variables with weak associations in one model that may otherwise be ignored by a procedure using only that model, such as the Bayesian lasso on the outcome model. Furthermore, both approaches show substantial improvements over BAC with $\omega = \infty$. SSCE aims to only adjust for covariates related to outcome or treatment assignment in order to minimize bias. Additionally, our proposed method, SSCE, very infrequently includes irrelevant covariates that are unrelated to outcome and treatment assignment, which can greatly reduce variability

and the MSE of the treatment effect estimator in finite samples compared to competing approaches such as BAC. On the other hand, BSSCE aims to reduce MSE of the treatment effect estimator by setting coefficients to zero for those covariates which if included in the outcome model would otherwise increase MSE of the treatment effect estimator. BSSCE provides an effective means of reducing MSE without having to data-adaptively choose a tuning parameter like ω in BAC. Furthermore, unlike BAC or other model averaging techniques, the proposed approaches performed well even when the number of covariates exceeded the sample size.

Another advantage of our approach is that it allows the variability of all parameters – including both the treatment effect and covariate inclusion probabilities – to be summarized from their posterior distributions. When using non-Bayesian methods such as the standard lasso or simultaneous variable selection technique proposed in Chapter 2, the variance of parameter estimators is often difficult to estimate; with many covariates or subjects, re-sampling methods can be computationally intensive as parameters must be re-estimated many times, and in some cases the bootstrap is not guaranteed to consistently estimate the relevant limiting distributions (Chatterjee and Lahiri, 2011).

A linear model was assumed throughout the, but the proposed method could be extended to more complicated models with covariate-covariate interactions, polynomial transformations of covariates, covariate-treatment interactions, or smoothers of various kinds (e.g., penalized splines expressed as mixed linear models). However, additional work needs to be done for such extensions as the ideal grouping structure for interactions and transformations is unclear.

3.7 Supplementary Materials

3.7.1 Supplement A: Gibbs sampler for SSCE and BSSCE

We now describe how to implement SSCE and BSSCE.

• First, set τ^2 to a large constant to prevent coefficient shrinkage bias (τ^2 is set to 1000 in all simulations and the application in the main manuscript).

The following conditional distributions are then used in the Gibbs samplers for SSCE and BSSCE (draw each parameter from its conditional distribution M times, where M is large):

$\bullet \sigma^2|\beta,\tau^2 \sim \text{IG}\left(\text{shape} = \frac{n}{2} + \frac{1}{2}\sum_{j=1}^{p} I(\beta_j \neq 0) + 0.1, \text{scale} = \frac{1}{2}\left(\mathbf{Y} - \mathbf{X}_{out}\beta\right)^T\left(\mathbf{Y} - \mathbf{X}_{out}\beta\right) + \frac{\beta^T\beta}{\tau^2} + 0.1\right)$,

where IG denotes the Inverse Gamma distribution,

•$\pi_0|\beta \sim \text{Beta}(1+p-\sum_{j=1}^{p} I(\beta_j \neq 0), 1+\sum_{j=1}^{p} I(\beta_j \neq 0))$.

• If $A_i = 0 : A_i^*|\gamma, \mathbf{X}_i \sim TN(\mathbf{X}_i\gamma, 1)^-$,

• If $A_i = 1 : A_i^*|\gamma, \mathbf{X}_i \sim TN(\mathbf{X}_i\gamma, 1)^+$,

where $TN(\mathbf{X}_i\gamma, 1)^-$ denotes the Truncated Normal distribution with support $(-\infty, 0)$ and $TN(\mathbf{X}_i\gamma, 1)^+$ denotes the Truncated Normal distribution with support $(0, \infty)$.

•$\beta_1|\sigma^2, \beta_{-1} \sim N\left(\frac{1}{n}\mathbf{1}^T\left(\mathbf{Y} - \mathbf{X}_{out,-1}\beta_{-1}\right), \frac{\sigma^2}{n}\right)$

•$\gamma_1|\gamma_{-1} \sim N\left(\frac{1}{n}\mathbf{1}^T\left(\mathbf{A}^* - \mathbf{X}_{trt,-1}\gamma_{-1}\right), \frac{1}{n}\right)$

•$\beta_A|\sigma^2, \beta_{-A} \sim N\left(\frac{1}{\sum_{i=1}^{n} A_i}\mathbf{A}^T\left(\mathbf{Y} - \mathbf{X}_{out,-A}\beta_{-A}\right), \frac{\sigma^2}{\sum_{i=1}^{n} A_i}\right)$

•$\beta_j|\beta_{-j}, \sigma^2, \tau^2, \pi_0 \sim (1-l_j)N\left(\mu_{1_j}, \frac{\sigma^2}{(n-1+1/\tau^2)}\right) + l_j\delta_0(\beta_j)$, j=1,. . . ,p,

•$\gamma_j|\gamma_{-j}, \sigma^2, \tau^2, \pi_0 \sim (1-l_j)N\left(\mu_{2j}, \frac{1}{(n-1+1/\tau^2)}\right) + l_j\delta_0(\gamma_j)$, j=1,. . . ,p, where

$\mu_{1_j} = \frac{1}{(n-1+1/\tau^2)}\mathbf{X}_{out,-j}^T(\mathbf{Y} - \mathbf{X}_{out,-j}\beta_{-j})$,

$\mu_{2_j} = \frac{1}{(n-1+1/\tau^2)}\mathbf{X}_{trt,-j}^T(\mathbf{A}^* - \mathbf{X}_{trt,-j}\beta_{-j})$, and

$l_j = \frac{\pi_0}{\pi_0 + (1-\pi_0)\left(\frac{\tau^{-2}}{n-1+1/\tau^2}\right)\exp\left(\frac{1}{2\sigma^2(n-1+1/\tau^2)^{1/2}}\mathbf{X}_{out,-j}^T(\mathbf{Y}-\mathbf{X}_{out,-j}\beta_{-j})^2 + \frac{1}{2(n-1+1/\tau^2)^{1/2}}\mathbf{X}_{trt,-j}^T(\mathbf{A}^*-\mathbf{X}_{trt,-j}\gamma_{-j})^2\right)}$.

• **If using BSSCE**:

• Calculate $Q_j = \text{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}) - \text{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}, X_j) + \text{Bias}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}, X_j)^2$ for $j = 1, \ldots, p$, where

$\text{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}) = \frac{1}{n}\frac{\sigma^2_{Y|A,\mathbf{X}_{-j}}}{\sigma^2_{A|\mathbf{X}_{-j}}}$,

$\text{Var}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}, X_j) = \frac{1}{n}\frac{\sigma^2_{Y|A,\mathbf{X}_{-j}} - \sigma^2_{X_j|\mathbf{X}_{-j},A}\,\beta_j^2}{\sigma^2_{A|\mathbf{X}} - \sigma^2_{X_j|\mathbf{X}_{-j}}\left(\frac{\gamma_j}{\sqrt{2\pi}}\right)^2}$, and

$\text{Bias}_{\hat{\beta}_A}(\beta_j, \gamma_j; \mathbf{X}_{-j}, X_j) = \beta_j\frac{\gamma_j}{\sqrt{2\pi}}\frac{\sigma^2_{X_j|\mathbf{X}_{-j}}}{\sigma^2_{A|\mathbf{X}_{-j}}}$.

• For $j = 1, \ldots, p$, if $Q_j < 0$ then set β_j and γ_j equal to zero.

3.7.2 Supplement B: A motivating example for SSCE

A simplified example is presented here to illustrate the properties and behavior of SSCE. Suppose we have outcome and treatment design matrices $\mathbf{M}_{out}$ and $\mathbf{M}_{trt}$ such that $\mathbf{M}_{out}^T\mathbf{M}_{out} = n\mathbf{I}_n$ and $\mathbf{M}_{trt}^T\mathbf{M}_{trt} = n\mathbf{I}_n$, and $\mathbf{M}_{out_i}^T\mathbf{M}_{out_j} = 0$, and $\mathbf{M}_{trt_i}^T\mathbf{M}_{trt_j} = 0$ for $i \neq j$, where $\mathbf{M}_{out_j}$ and $\mathbf{M}_{trt_j}$ denote the jth columns in $\mathbf{M}_{out}$ and $\mathbf{M}_{trt}$, respectively. Note that this means there are no columns for intercepts in $\mathbf{M}_{out}$ and $\mathbf{M}_{trt}$ and also no column for the treatment effect in $\mathbf{M}_{out}$ (i.e., $\beta_A = 0$). When BSSL is used on only $\mathbf{M}_{out}$, the posterior probability

that β_j is zero is

$$P(\beta_j = 0|\mathbf{Y}, \mathbf{M}_{out}) = \frac{\pi_0}{\pi_0 + (1-\pi_0)(1+n\tau_j^2)^{-1/2} \exp\left\{\frac{n\tau_j^2}{2\sigma^2(1+n\tau_j^2)} n|\hat{\beta}_j^{ls}|\right\}}, \tag{3.3}$$

where $\hat{\beta}_j^{ls}$ is the least squares estimator of β_j under a full model. Equation (3.3) is plotted in black in Figure 3.1 of the main manuscript (for fixed τ_j^2); the colored lines in Figure 3.1 of the main manuscript show the posterior probability that β_j is zero under SSCE (for identical values of the π_0, n, σ^2, and τ_j^2), which is

$$P(\beta_j = 0|\mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt}) = \frac{\pi_0}{\pi_0 + (1-\pi_0)(1+n\tau_j^2)^{-1} \exp\left\{\frac{n\tau_j^2}{2\sigma^2(1+n\tau_j^2)} n\sqrt{\hat{\beta}_j^{ls^2} + \tilde{\gamma}_j^2}\right\}}, \tag{3.4}$$

where $\tilde{\gamma}_j$ is the maximum likelihood estimator of γ_j under a full model.

The posterior distribution of β_j for fixed τ_j^2 under design matrices $\mathbf{M}_{out}$ and $\mathbf{M}_{trt}$ is a spike and slab distribution,

$$\begin{aligned}\beta_j|\mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt} \sim\ & P(\beta_j = 0|\mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt})\delta_0(\beta_j) \\ & + (1 - P(\beta_j = 0|\mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt}))\mathrm{N}\left(\hat{\beta}_j^{ls}(1-B), \frac{\sigma^2}{n}(1-B)\right),\end{aligned} \tag{3.5}$$

where $P(\beta_j = 0|\mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt})$ is given in (3.4) and $B = \frac{1}{1+n\tau_j^2}$.

3.7.3 Supplement C: Covariate inclusion criteria for SSCE and BSSCE

Now we only assume that the design matrices for the outcome and treatment models, $\mathbf{X}_{out}$ and $\mathbf{X}_{trt}$, have standardized covariates (as in Chapter 3.2.2 of the main manuscript). The conditional probability that β_j (and γ_j) is zero under SSCE is (with $\mathbf{Z}$ and $\mathbf{O}$ defined as in Chapter 3.3.1 of the main manuscript)

$$P((\beta_j, \gamma_j)^T = 0|\mathbf{Z}, \mathbf{O}, A, \sigma^2, \pi_0, \tau_j^2, \beta_0, \beta_A, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p, \gamma_0, \gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p) =$$

$$\frac{\pi_0}{\pi_0 + (1-\pi_0)\sigma\tau_j^{-2}\left(n - 1 + \frac{1}{\tau_j^2}\right)^{-1} \exp\left\{\frac{\left(n-1+\frac{1}{\tau_j^2}\right)^{-1}}{2}\left(\frac{a^2}{\sigma^2} + b^2\right)\right\}},$$

where

$$a = \mathbf{X}_j^T \left(\mathbf{Y} - \beta_0 - \beta_A \mathbf{A} - \sum_{k \neq j}^{p} \beta_k \mathbf{X}_k \right) \text{ and } b = \mathbf{X}_j^T \left(\mathbf{A}^* - \gamma_0 - \sum_{k \neq j}^{p} \gamma_k \mathbf{X}_k \right),$$

with $\mathbf{X}_j$ denoting the column in $\mathbf{X}$ for the jth covariate. Note that a and b are proportional to the correlation between the jth covariate and the residual vectors without the jth covariate in the outcome and treatment models, respectively.

Chapter 4

A matching-based approach to assessing treatment effect heterogeneity using conditional inference trees

4.1 Introduction

Randomized experiments are conducted every day to estimate the causal effect of a particular treatment or intervention on some outcome. There are currently a quarter of a million interventional studies that are registered and have results posted on ClinicalTrials.gov. Web-facing companies such as Amazon, Facebook, Google, and Netflix use online controlled experiments to guide product development and accelerate innovation; at Microsoft's Bing, over 200 concurrent experiments are now running on any given day (Kohavi et al., 2013). Results from a randomized experiment may sometimes suggest that a treatment or intervention is particularly beneficial (or harmful) to some subgroup. When there are differential effects of treatment, we say there is treatment effect heterogeneity.

Policy and medical decisions are often informed by the results of randomized studies, so correctly characterizing treatment effect heterogeneity is important. For example, BiDil is the first drug to be FDA-approved for a single racial group because study results suggest it is beneficial for African-Americans with congestive heart failure. However, if BiDil is not

truly beneficial for African-Americans with congestive heart failure (i.e., if a Type I error was made), then production and marketing of the drug would prove to be very wasteful and the potential side effects of BiDil would likely cause harm with no benefit for patients who have received the drug.

Traditional approaches to characterizing treatment effect heterogeneity have centered around regression modeling with interaction terms between the treatment/intervention indicator and covariates. The magnitude and statistical significance of the interaction term can be used to assess the degree of treatment effect heterogeneity. However, when there are multiple interactions, regression models can quickly become difficult to interpret. Also, conclusions about the presence of interaction depend on the way in which interaction terms are specified, so that important subgroups may not be identified. Lastly, considering multiple interactions can inflate Type I error, but it is common not to adjust the significance level for multiple testing, which leads to anti-conservative inference (Green and Kern, 2012).

Recently, there has been increasing interest in developing techniques to estimate heterogeneous treatment effects using more flexible models that can "automatically" detect subgroups of interest. Athey and Imbens (2016) developed Causal Tree, which uses a single regression tree to recursively partition the data into homogeneous subgroups that have similar treatment effects and a similar subset of covariate values. Wager and Athey (2018) extend Causal Tree to Causal Forests, which averages treatment effect estimates over many Causal Trees. Green and Kern (2012) propose another method that uses multiple regression trees, called Bayesian Adaptive Regression Trees (BART), which automatically detects nonlinear relationships and interactions to describe treatment effect heterogeneity. While many of these approaches provide flexible subgroup identification, they typically do not address the issue of inflation of Type I error.

In this chapter, we propose a novel method, called Treatment Effect Heterogeneity Trees (TEHTrees), for characterizing treatment effect heterogeneity. TEHTrees combines matching with conditional inference trees (Hothorn, Hornik, and Zeileis, 2006). One of its main virtues is that, by employing formal hypothesis testing procedures in constructing the tree, TEHTrees preserves the Type I error rate. In simulation studies comparing TEHTrees and Causal Tree, the Type I error rate is below 0.05 in all considered scenarios using TEHTrees, but is greater than 0.15 in all scenarios using Causal Tree (and far greater in some cases). Though the power (defined as the probability of splitting on the variable

with true heterogeneous treatment effects) is slightly larger using Causal Tree compared to TEHTrees in our simulations with continuous covariates, the power with binary covariates is actually greater using TEHTrees compared to that of Causal Tree.

4.2 Method

4.2.1 Notation and Terminology

Let $Y = \{Y_1, \ldots, Y_k, Y_{k+1}, \ldots, Y_n\}$ be the observed outcomes for k subjects randomized to the control group and $n - k$ subjects randomized to the treatment group. Throughout we assume Y is a continuous outcome although the method could be generalized to binary or categorical outcomes. An accompanying $n \times m$- dimensional matrix is denoted as $\mathbf{X} = \{X_1, \ldots, X_m\}$, where X_j denotes the jth predictor/covariate, and contains the m predictors for each of the n subjects. The treatment assignment for all subjects is denoted as $Z = \{Z_1, \ldots, Z_n\}$.

4.2.2 Treatment Effects and Matching

Let Y_{1i} be the potential outcome of subject i if assigned to treatment and Y_{0i} be the potential outcome of subject i if assigned to control. The measure of treatment effect heterogeneity for a subject with covariates X_i is $E(Y_{1i} - Y_{0i}|X_i)$. If we observed Y_{1i} and Y_{0i} for all i, then to estimate $E(Y_{1i} - Y_{0i}|X_i)$ we could simply construct a regression tree on the differences of $Y_{1i} - Y_{0i}$ to get an overall estimate of treatment effect heterogeneity within the population. However, this is a counterfactual conditional statement because we do not observe both Y_{1i} and Y_{0i} for any i. This is the fundamental difficulty with causal inference: for each individual, we can only observe the outcome under treatment or the outcome under control because each individual at a particular time will only receive the treatment or control.

One approach to estimating individual-level effects is via matching. In general, matching is used to find a control subject j that is equivalent or similar across all covariates for each treatment subject i. Assuming $X_i = X_j$, we can then use $E(Y_i - Y_j|X_i)$ to estimate $E(Y_{1i} - Y_{0i}|X_i)$ for all i. That is, we can construct a regression tree (see next subchapter for more on regression trees) on $Y_i - Y_j$, with X_i as the regression covariates in the tree, to estimate $E(Y_i - Y_j|X_i)$, which should approximate $E(Y_{1i} - Y_{0i}|X_i)$. The resulting regression tree partitions the covariate space into homogeneous subpopulations that have similar covariate

values and estimates of $E(Y_{1i} - Y_{0i}|X_i)$. The subpopulations formed from the regression tree are defined hierarchically via a sequence of binary partitions of the covariate space. If the predicted outcome for a subpopulation is much greater than that of another subpopulation, then there is evidence that the treatment effect is heterogeneous in the study population (i.e., the treatment is more beneficial or harmful to some subpopulation).

This approach requires implementation of an algorithm to match treated individuals with controls. Stuart (2010) provides a comprehensive overview of past and current research into matching methods and guidance on their use. She defines matching broadly as "any method that aims to equate (or balance) the distribution of covariates in the treated and control groups." Several common methods include 1:1 matching, weighting, and subclassification. Theoretical basis for these procedures was developed in the 1970s beginning with papers by Cochran and Rubin (1973) and Rubin (1973a) for situations with one covariate. Matching on a single covariate is much simpler than matching on multiple covariates. As the number of covariates increases, so does the difficulty in finding matches with close or exact values of all covariates. Matching on the propensity score, defined as the probability of receiving the treatment given the observed covariates, simplifies matching by not requiring close or exact matches across all of the individual variables (Rosenbaum and Rubin, 1983; Stuart, Lee, and Leacy , 2013). It accomplishes this by collapsing the covariates into a single measure, summarizing their joint association with treatment conditions (Hansen, 2008). Traditional propensity score methods are conducted without use of the outcome variable, so the same propensity score procedure can be used for multiple outcomes. However, this prioritizes variables by their importance in predicting treatment assignment, not outcome. Thus, variables that are strongly related to treatment assignment, but unrelated to the outcome of interest may have excessive influence in the propensity score resulting in decreased precision and increased bias in the treatment effect estimator (Stuart, Lee, and Leacy , 2013).

A lesser known, but equally effective, measure that collapses an $n \times m$ matrix of covariates into an n-dimensional vector is the prognostic score, formalized by Hansen (2008). We call $\phi(X)$ a prognostic score if $\phi(X)$ is sufficient for Y_0, in the sense that Y_0 is independent of X given $\phi(X)$. This may have several interpretations depending on the distribution of the potential responses to control, Y_0. Should $Y_0|X$ follow a generalized linear model, then the linear predictor of Y_0 given X is a prognostic score, as is the scalar $E(Y_0|X)$ (Hansen,

2008). In this scenario, a prognostic score is estimated by fitting a model (we use the Super Learner by van der Laan, Polley, and Hubbard (2007) in our simulations) of the outcome in the control group and then using that model to obtain predictions of the outcome under the control condition for all individuals (Stuart, Lee, and Leacy , 2013). The advantage to using the prognostic score is that it reflects the relative importance of each covariate in predicting the outcome. However, diagnostics for prognostic balance are inherently incomplete in that they address only balance among controls (Hansen, 2008). Fitting prognostic scores using both treatment and control observations could also introduce bias if the pattern of the treatment effect is not known a priori.

4.2.3 Conditional Inference Trees

As a solution to the downfalls of the standard approaches for developing clusters of individuals with shared characteristics, Morgan and Sonquist (1963) introduced a methodology they called decision trees. Decision trees begin with one root node, containing the entire sample, sometimes called the parent node. They then examine all possible independent variables and select the one such that, if the sample is split according to that variable, the resulting groups are most different with respect to the dependent variable according to a predetermined criterion. The parent node then branches into two mutually-exclusive child nodes according to the independent variable and split point that was selected. Within each of the two child nodes, the tree-growing methodology repeats until a stopping rule is satisfied. The final, mutually-exclusive and exhaustive subgroups of the population are called terminal nodes. Since its development, the decision tree methodology has been applied in many areas of research including economics (Fan, Ong, and Koh, 2006), political science (Feller and Holmes, 2009; Green and Kern, 2012), and life science (Gass et al., 2014; Pouliakis et al., 2014; Faries et al., 2013).

Breiman et al. (1984) introduced a popular decision tree methodology called Classification and Regression Tree (CART) analysis during the mid-1980s, which uses mean squared error (MSE) of the outcomes as its splitting criterion. Despite its widespread popularity, CART has several drawbacks, including a tendency to split the sample on variables that have more potential splitting points (e.g. samples are more likely to be split on continuous variables than binary when each has a similar association with the outcome). Also, CART is predisposed to overfit the model because it is greedy when searching for optimal splits.

Thus, it is often necessary to consider more or less aggressive stopping rules, or prune, in order to obtain a tree that predicts well without overfitting in validation datasets.

One alternative decision tree method is the Conditional Inference Tree (CIT), proposed by Hothorn, Hornik, and Zeileis (2006). CITs are different from CARTs in that the processes of determining the variable to split on and determining the optimal split of the selected variable occur sequentially instead of simultaneously. As our method relies on CITs, we will provide a more in depth explanation of this algorithm in the following paragraphs.

For each parent node, the association on the p-value scale between the outcome and each covariate is used to select the most optimal variable on which to split. P-values are used in place of test statistics because the latter depend on the measurement scale and number of possible splits of the covariates. The global null hypothesis of independence is tested by considering all m partial null hypotheses of independence between the outcome and each covariate. The global null hypothesis is rejected if the minimum computed p-value for the m partial null hypotheses is less than a pre-specified level of significance. Multiple comparison adjustment, such as the Bonferroni adjustment, is applied to the partial null hypothesis tests. Finally, the partial null hypothesis that results in the smallest p-value will indicate the covariate that is most strongly associated with the outcome. In the case where no partial tests are rejected, the node is deemed to be a terminal node and no split is made.

After a covariate (say X_l) has been selected for splitting, the second step of the CIT algorithm is to find the split point s that maximizes the discrepancy between the two samples of outcomes that are obtained within the node based on if $X_l < s$. That is, if we temporarily let A denote the set of indices that belong to the node under consider, for each possible split point s, one can calculate a test statistic T_s that measures the discrepancy between $\{Y_i|X_{li} < s, i \in A\}$ and $\{Y_i|X_{li} \geq s, i \in A\}$, and the desired split point maximizes $|T_s|$. Some restrictions on the chosen split point (i.e., minimum or maximum values) may be used to ensure the samples $\{Y_i|X_{li} < s, i \in A\}$ or $\{Y_i|X_{li} < s, i \in A\}$ do not have too few observations.

This two-step process for determining splits provides both advantages and drawbacks over the CART algorithm. Venkatasubramaniam et al. (2017) showed that CART may provide slightly higher predictive accuracy and is less sensitive to sample size compared to CIT. However, nodes formed through the CIT algorithm have a quantifiable relationship between covariates and outcome amongst the individuals belonging to that node. Also,

there is less worry about overfitting models using CIT so pruning is rarely necessary. Finally, all covariates, assuming common association with the outcome, are similarly likely to be selected as the splitting candidate no matter the number of potential split points.

4.2.4 TEHTrees

Characterizing treatment effect heterogeneity would be difficult using a CIT with the observed outcomes (Y) and covariates (X) because terminal nodes could contain no treated (or untreated) subjects, which would complicate treatment effect estimation and obscure intuition about which subgroups experience beneficial (or harmful) treatment effects. Additionally, since splits do not necessarily involve the treatment assignment, there would be a greater chance that the tree would stop growing before important heterogeneous treatment effects are discovered, especially when using a conservative method for multiple comparison adjustment (like Bonferroni). The algorithm for characterizing treatment effect heterogeneity using a CIT is:

1. Split the sample in half.

   Split the sample so that approximately half of the treated and half of the controls are in one sample (call this the training sample), and the remaining subjects are in the other sample (call this the estimation sample). The honest estimation technique of Causal Tree also uses separate samples for tree-building and estimation.

2. Match each treated subject to a control subject (with replacement).

   Form a set of matched pairs from the training the data, such that each pair contains one treated and one control subject which are closely matched based on their prognostic score. In particular, the Super Learner (van der Laan, Polley, and Hubbard, 2007) is used on the controls in the training sample only to model the outcome as a function of the covariates (i.e., not including indicator for treatment assignment), and this model is used to obtain the prognostic score (i.e., the estimated conditional mean of the outcome given the observed covariates) for each individual in the training sample. Each treated subject in the training sample is matched with the control that has the prognostic score nearest (based on absolute distance) the prognostic score of the treated subject; ties are broken randomly. Note that a control in the training sample may be matched to multiple treated subjects, while it's also possible that a

control is not matched to any treated subject.

3. Calculate the within-pair differences in the outcome.

   For each pair (there are as many pairs as treated subjects in the training sample), calculate the within-pair difference in the outcome. We will use the within-pair differences as outcomes in a CIT.

4. Build the tree.

   Perform the partitioning within the CIT by repeating the following at each parent node:

   4a. Determine the splitting variable.

   The p-values that determine whether to split the node (i.e., the p-values in the first step of the iterative algorithm of the CIT described in Chapter 4.2.3) are derived using a linear mixed effects model. Since matching is performed with replacement (i.e., the same control subject could be matched to two or more different treated subjects), the within-pair differences in outcomes are correlated even though the original observations are independent. This results in underestimated variances within a linear regression model, and hence inflated Type I error from the standard CIT algorithm. As an alternative, we use a linear mixed effects model to account for the correlated outcomes. Particularly, the within-pairs difference in the outcomes are modeled as the response in a linear mixed effects model with a random intercept denoting the control that is used for each matched pair. Each model contains a single covariate as a fixed effect. For each parent node, we fit a separate model for each covariate, so that there are m total models (we fit models using the `R` package `nlme`). For each model, a p-value is calculated using the Wald test corresponding the the null hypothesis that the fixed effect is zero for the covariate in the model (a different model-based test that measures the association between the within-pairs difference and each covariate may also be used). If the smallest p-value (there are m total p-values) is greater than or equal to $\frac{\alpha}{m}$, where α is the level of significance, then the node is a terminal node (i.e., it is not split). Otherwise, the covariate corresponding to the smallest p-value is chosen for splitting the node, and the split point of this covariate is described in the next sub-step.

   4b. Determine the split point.

To determine the split point of a continuous covariate X_l, first consider a finite number of split points, $s_1, \ldots, s_r$ (we consider equally spaced quantiles in our simulations). The points s_1 or s_r can be restricted so that there are a minimum number of treated and control observations in each of the child nodes in the estimation sample; note that at least one treated and one control observation in the estimation sample are needed in each terminal node for the estimation step below (we use a minimum of 10 treated and 10 controls per node in the estimation sample). At the jth considered split point, we use a linear mixed effects model with the same responses and random effect as the previous sub-step, but with the covariate replaced with $I(X_l < s_j)$; the split point s_j that yields the largest absolute coefficient for $I(X_l < s_j)$ is chosen as the split point for the node.

5. Estimate treatment effects.

   Use the estimation sample to estimate the desired causal treatment effect in each terminal node. We use the estimation technique of Causal Tree to estimate the treatment effect in each node. The estimated treatment effect in a particular terminal node is the mean of the outcomes of those treated in the estimation sample that belong to the node (according to their covariate values) minus the mean of the outcomes of the untreated subjects in the estimation sample that belong to the node. If the average effect on the treated is desired, one can instead use as an estimate the mean of the outcomes of the treated subjects in the estimation sample in each terminal node.

4.3 Simulation Study

We conducted simulations using R 3.2.2 (R Core Team, 2015) to test TEHTrees and compare its performance to Causal Tree (Athey and Imbens, 2016), which extends the CART algorithm to allow causal treatment effect estimation. Matching was conducted using the Matching package (Sekhon , 2011), and all linear mixed models were fit using the nlme package (Pinheiro et al., 2015). The Super Learner was used to estimate the prognostic score and used the sample mean, a linear model (with and without interaction terms), a generalized additive model, a Random Forest, stepwise regression (with and without interaction terms), and "polymars" (multivariate adaptive polynomial spline regression) as base learners. The tree generating function TEHTrees had arguments specifying a minimum of

20 observations for a node to be considered for splitting and a p-value threshold of 0.05 for determining a candidate for splitting. Potential split points were tested at every 0.05 quantile of continuous candidate covariates.

4.3.1 Data Generation

The sample size N is varied in each scenario and the treatment variable Z was generated such that $N/2$ subjects received treatment ($Z = 1$) and the $N/2$ subjects received control ($Z = 0$). Continuous covariates were generated from multivariate normal distributions with mean zero, unit variance, and varying pairwise correlations. Binary covariates were generated as independent Binomial$(1, 0.5)$. Continuous outcomes were generated as independent N$(\mu, 1)$ with linear predictor μ of the form:

$$g(\mu) = \alpha + \theta Z + \beta \mathbf{X} + \gamma I(X_1 > 0)Z.$$

We set α and θ to 0.8, β to $(1.0, 0.8, 0.6, 0.4, 0.2)^T$ for the first five covariates (0 otherwise), and let γ equal either 0 (meaning there is no treatment effect heterogeneity) or 3 (meaning there is treatment effect heterogeneity). Results are based on 1,000 simulations per scenario. The total number of covariates, type of covariates, and pairwise correlation between covariates were varied in addition to the sample size.

4.3.2 Type I Error

In the context of this study, we define the Type I error to be any tree that results in more than one terminal node when there is no treatment effect heterogeneity (i.e., when $\gamma = 0$). Table 4.1 displays the Type I error of TEHTrees and Causal Tree under a variety of scenarios. The Type I error of TEHTrees is less than the desired 0.05 level in all 23 scenarios. The Type I error of Causal Tree, on the other hand, is greater than 0.15 in every scenario. As the sample size increases, the Type I error of Causal Tree increases and even almost reaches 1.0 with 5 continuous covariates and $N = 2000$; the Type I error of TEHTrees shows no pattern as the sample size increases.

Table 4.1: Type I Error rate of TEHTrees and Causal Tree when there is no treatment effect heterogeneity (i.e., when $\gamma = 0$). A Type I error occurs if the tree has more than one terminal node (i.e., the tree splits on any variable). N is sample size, m is number of covariates, and ρ is pairwise correlation.

Covariate Type	N	m	ρ	Type I Error Rate	
				TEHTrees	Causal Tree
Binary	100	5	0.0	0.020	0.630
Binary	200	5	0.0	0.046	0.998
Binary	500	5	0.0	0.047	1.000
Binary	1000	5	0.0	0.041	1.000
Binary	2000	5	0.0	0.039	1.000
Continuous	100	5	0.0	0.025	0.303
Continuous	200	5	0.0	0.019	0.646
Continuous	500	5	0.0	0.025	0.936
Continuous	1000	5	0.0	0.027	0.991
Continuous	2000	5	0.0	0.026	1.000
Continuous	100	10	0.0	0.028	0.205
Continuous	200	10	0.0	0.016	0.416
Continuous	500	10	0.0	0.017	0.821
Continuous	1000	10	0.0	0.010	0.957
Continuous	2000	10	0.0	0.018	0.996
Continuous	200	5	0.2	0.030	0.687
Continuous	200	5	0.4	0.026	0.767
Continuous	200	5	0.6	0.026	0.825
Continuous	200	5	0.8	0.050	0.851
Continuous	500	5	0.2	0.023	0.970
Continuous	500	5	0.4	0.026	0.991
Continuous	500	5	0.6	0.030	0.994
Continuous	500	5	0.8	0.045	0.998

4.3.3 Power

The power is defined as the probability that the tree produces a split on X_1 (the variable with heterogeneity in treatment effects) when there is treatment effect heterogeneity (i.e., when $\gamma = 3$). Table 4.2 shows the power of TEHTrees and Causal Tree under the considered scenarios. With binary covariates, TEHTrees displays greater power than Causal Tree for all

Table 4.2: Power of TEHTrees and Causal Tree. Power is defined to be the probability of the tree making a split on X_1 when there is treatment effect heterogeneity (i.e., when $\gamma = 3$). N is sample size, m is number of covariates, and ρ is pairwise correlation.

Covariate Type	N	m	ρ	Power	
				TEHTrees	Causal Tree
Binary	100	5	0.0	0.62	0.48
Binary	200	5	0.0	1.00	0.69
Binary	500	5	0.0	1.00	0.84
Binary	1000	5	0.0	1.00	0.91
Binary	2000	5	0.0	1.00	0.94
Continuous	100	5	0.0	0.29	0.71
Continuous	200	5	0.0	0.63	0.94
Continuous	500	5	0.0	0.93	0.99
Continuous	1000	5	0.0	1.00	1.00
Continuous	2000	5	0.0	1.00	1.00
Continuous	100	10	0.0	0.18	0.54
Continuous	200	10	0.0	0.41	0.89
Continuous	500	10	0.0	0.81	1.00
Continuous	1000	10	0.0	0.96	1.00
Continuous	2000	10	0.0	1.00	1.00
Continuous	200	5	0.2	0.68	0.93
Continuous	200	5	0.4	0.72	0.89
Continuous	200	5	0.6	0.73	0.88
Continuous	200	5	0.8	0.66	0.83
Continuous	500	5	0.2	0.97	0.98
Continuous	500	5	0.4	0.98	0.98
Continuous	500	5	0.6	0.98	0.95
Continuous	500	5	0.8	0.95	0.94

sample sizes. With continuous covariates, the power is greater for Causal Tree with smaller sample sizes compared to TEHTrees, but the power is similar between the two methods with larger sample sizes.

Table 4.3: Characteristics of the trees built by TEHTrees (TT) and Causal Tree (CT) when there is treatment effect heterogeneity (i.e., when $\gamma = 3$), including the median and mean of the first split point on X_1 (the variable with heterogeneous treatment effects), along with the proportion of those split points that are within the middle 5% of a standard normal distribution (i.e., proportion in $I = (-0.063, 0.063)$), since the true split point is zero. The average number of terminal nodes when a split is made is also shown. N is sample size, m is number of covariates, and ρ is pairwise correlation.

Covariate type	N	m	ρ	Median split point		Mean split point		Prop. splits in I		# terminal nodes	
				TT	CT	TT	CT	TT	CT	TT	CT
Binary	100	5	0.0	-	-	-	-	-	-	2.00	1.75
Binary	200	5	0.0	-	-	-	-	-	-	2.06	2.21
Binary	500	5	0.0	-	-	-	-	-	-	2.11	2.56
Binary	1000	5	0.0	-	-	-	-	-	-	2.11	2.99
Binary	2000	5	0.0	-	-	-	-	-	-	2.11	3.37
Continuous	100	5	0.0	-0.01	-0.01	-0.01	-0.01	0.40	0.25	2.00	2.47
Continuous	200	5	0.0	0.00	0.00	0.00	0.00	0.53	0.39	2.10	3.90
Continuous	500	5	0.0	0.00	0.00	0.00	0.00	0.71	0.75	2.21	7.41
Continuous	1000	5	0.0	0.00	-0.00	0.00	-0.00	0.82	0.95	2.46	15.23
Continuous	2000	5	0.0	0.00	-0.00	0.00	-0.00	0.87	0.99	2.91	31.35
Continuous	100	10	0.0	-0.01	-0.03	-0.03	-0.02	0.40	0.24	2.00	2.29
Continuous	200	10	0.0	0.00	0.01	0.02	0.01	0.55	0.39	2.08	3.62
Continuous	500	10	0.0	0.00	0.00	0.00	0.00	0.73	0.64	2.16	4.39
Continuous	1000	10	0.0	0.00	0.00	0.00	0.00	0.82	0.86	2.34	4.56
Continuous	2000	10	0.0	0.00	0.00	0.00	-0.00	0.87	0.98	2.96	6.63
Continuous	200	5	0.2	0.00	0.00	0.00	0.00	0.51	0.40	2.13	4.43
Continuous	200	5	0.4	0.00	-0.00	0.00	-0.00	0.53	0.44	2.14	4.58
Continuous	200	5	0.6	0.00	0.00	-0.01	-0.00	0.49	0.43	2.16	4.76
Continuous	200	5	0.8	0.00	0.00	0.01	-0.01	0.52	0.44	2.16	4.54
Continuous	500	5	0.2	0.00	-0.00	0.00	-0.00	0.72	0.79	2.24	9.87
Continuous	500	5	0.4	0.00	-0.00	-0.01	-0.00	0.74	0.78	2.21	10.07
Continuous	500	5	0.6	-0.01	-0.00	0.00	-0.00	0.73	0.78	2.20	9.56
Continuous	500	5	0.8	-0.01	-0.00	-0.01	-0.01	0.73	0.75	2.24	9.10

4.3.4 Tree Characteristics Under Treatment Effect Heterogeneity

Table 4.3 provides characteristics of the trees built by TEHTrees and Causal Tree when there is treatment effect heterogeneity (i.e., when $\gamma = 3$), including the median, mean, and standard deviation of the set of points denoting the first split on X_1 (the variable with heterogeneous treatment effects), along with the proportion that are within the middle 5% of a standard normal distribution (i.e., proportion in $(-0.063, 0.063)$) as the true split point is zero. We can see from Table 4.3 that the mean and median split points are near zero for both TEHTrees and Causal Tree, but the variability is greater for TEHTrees compared to

Causal Tree. The proportion of split points in $(-0.063, 0.063)$ is greater in smaller sample sizes, but smaller in larger samples for TEHTrees compared to Causal Tree.

Table 4.3 also shows the average number of terminal nodes per tree in which a split is made for TEHTrees and Causal Tree when there is treatment effect heterogeneity. The average number of terminal nodes is near 2 when a split is made using TEHTrees, while this number is much greater using Causal Tree; with $N = 2000$ independent continuous covariates, there is an average of 31.35 terminal nodes using Causal Tree.

4.3.5 Treatment Effect Estimation

To evaluate the estimation properties of TEHTrees and Causal Tree when there was treatment effect heterogeneity, we predicted the average treatment effect for all subjects i in the estimation sample such that $X_{1i} > 0$ (the subjects experiencing the most positive treatment effect). The bias, standard deviation, and MSE of the treatment effect estimates are shown in Table 4.4. With binary covariates, the magnitude of the bias and the MSE are much smaller using TEHTrees compared to Causal Tree, and even with the largest considered sample size ($N = 2000$), the MSE using TEHTrees is about 15 times smaller than that of Causal Tree. With continuous covariates, the estimation performance is better using Causal Tree compared to TEHTrees, but the differences in results between the methods are smaller than with binary covariates; with five independent continuous covariates and $N = 2000$ subjects, the MSE using Causal Tree is only about 1.5 times smaller than that of TEHTrees.

4.4 Discussion

Randomized studies are often used to make medical decisions and inform policy. Sometimes the results of randomized studies suggest that treatment effects are hetergeneous in a population, which may be used to pursue interventions or market treatments that target specific subgroups of the population. Recently, numerous methods have been proposed that aim to discover treatment effect heterogeneity, including Causal Tree by Athey and Imbens (2016) (which was later extended to Causal Forests by Wager and Athey (2018)). While many of these approaches provide flexible subgroup identification, they typically do not address the

Table 4.4: Bias, standard deviation (SD), and MSE of the estimated average treatment effect for the subjects in the estimation sample that have $X_1 > 0$ (i.e., the subjects in the estimation sample with the greatest treatment effect) using TEHTrees (TT) and Causal Tree (CT). N is sample size, m is number of covariates, and ρ is pairwise correlation.

Covariate type	N	m	ρ	Bias		SD		MSE	
				TT	CT	TT	CT	TT	CT
Binary	100	5	0.0	-0.57	-0.81	0.88	0.87	1.09	1.40
Binary	200	5	0.0	0.02	-0.51	0.32	0.76	0.10	0.84
Binary	500	5	0.0	0.01	-0.27	0.20	0.60	0.04	0.43
Binary	1000	5	0.0	0.00	-0.15	0.15	0.45	0.02	0.23
Binary	2000	5	0.0	0.00	-0.09	0.10	0.37	0.01	0.15
Continuous	100	5	0.0	-1.15	-0.75	0.90	0.86	2.13	1.29
Continuous	200	5	0.0	-0.71	-0.35	0.80	0.57	1.14	0.45
Continuous	500	5	0.0	-0.20	-0.13	0.46	0.33	0.26	0.12
Continuous	1000	5	0.0	-0.09	-0.05	0.23	0.20	0.06	0.04
Continuous	2000	5	0.0	-0.08	-0.02	0.15	0.13	0.03	0.02
Continuous	100	10	0.0	-1.30	-0.93	0.83	0.88	2.37	1.64
Continuous	200	10	0.0	-1.01	-0.44	0.80	0.63	1.65	0.58
Continuous	500	10	0.0	-0.37	-0.13	0.64	0.32	0.54	0.12
Continuous	1000	10	0.0	-0.14	-0.07	0.36	0.21	0.15	0.05
Continuous	2000	10	0.0	-0.08	-0.04	0.17	0.15	0.03	0.02
Continuous	200	5	0.2	-0.65	-0.40	0.82	0.64	1.10	0.57
Continuous	200	5	0.4	-0.56	-0.42	0.81	0.68	0.96	0.63
Continuous	200	5	0.6	-0.57	-0.43	0.82	0.68	1.10	0.64
Continuous	200	5	0.8	-0.59	-0.46	0.80	0.73	0.99	0.74
Continuous	500	5	0.2	-0.14	-0.14	0.43	0.38	0.20	0.16
Continuous	500	5	0.4	-0.13	-0.13	0.40	0.41	0.18	0.18
Continuous	500	5	0.6	-0.13	-0.17	0.40	0.45	0.18	0.23
Continuous	500	5	0.8	-0.16	-0.18	0.44	0.44	0.22	0.23

issue of preserving Type I error rate. We propose TEHTrees, a new method for characterizing treatment effect heterogeneity that preserves the Type I error rate. TEHTrees uses decision trees to characterize treatment effect heterogeneity by utilizing matching within a conditional inference tree algorithm.

In simulation experiments, the Type I error rate using TEHTrees was below 0.05 (the pre-specified significance level) in all 23 considered scenarios, while the Type I error using

Causal Tree was at least 0.15 in every scenario and even greater than 0.9 in many cases. Though Causal Tree has slightly greater power when there are continuous covariates, compared to TEHTrees, the power when using TEHTrees is actually greater than the power when using Causal Tree with binary covariates. When there is treatment effect heterogeneity, Causal Tree also tends to grow larger trees with more terminal nodes compared to TEHTrees, particularly with larger sample sizes. This makes it more difficult to infer the characteristics of groups that truly respond more (or less) to treatment when using Causal Tree compared to TEHTrees.

With binary covariates, treatment effect estimation was improved in our simulations when using TEHTrees compared to Causal Tree. However, Causal Tree displayed slightly better estimation properties than TEHTrees with continuous covariates, which was most likely due to greater variability in split points with TEHTrees. We conjecture that the variability in split points is larger with TEHTrees than with Causal Tree due to bias in the matching estimator. Decreasing bias in the matching estimator, or using an alternative approach to estimating the outcomes that are used as inputs in the conditional inference tree of TEHTrees, may therefore improve estimation of treatment effects with TEHTrees when there are continuous covariates.

TEHTrees is a flexible algorithm that allows for numerous modifications. A different matching algorithm can be implemented, an alternative prognostic score model can be fitted, other criterion can be used to find the splitting variable or its split point, and another estimation technique could be contrived and executed. The Bonferroni correction method that is used to find the splitting variable in TEHTrees (in Step 4a) is likely too conservative to detect small treatment effects when there are a large number of covariates in the study. Alternative multiple comparison adjustment methods should be explored in the case when there are many covariates. Controlling the false discovery rate may also be a desirable alternative to multiple comparison adjustment methods when there are numerous covariates. TEHTrees also does not account for the variability in the matching estimator, so an extra step may be needed to control for the inflation of Type I error that might occur in situations when good matches are difficult to obtain.

Though many simulation scenarios were considered in this chapter, a larger variety of scenarios need to be considered. The prognostic score model, for example, was quite simple. Future research will increase the complexity of the prognostic score model and

will also explore alternative effect sizes and treatment effect patterns (such as continuous interactions with treatment). Additionally, though we assume treatment is randomized (as in a clinical trial) throughout this chapter, TEHTrees could be extended for use with observational data; however, additional assumptions and steps in the algorithm would be required to achieve balance in the covariates, so this extension is left for future research.

Chapter 5

Conclusion and Discussion

Doubly robust estimation of the average causal treatment effect requires working models for both the outcome and treatment given possible confounders. When the number of possible confounders is large it is natural to consider some form of variable selection for the outcome and treatment models. GLiDeR, which is proposed in Chapter 2, uses an adaptive group lasso approach to perform coefficient regularization and estimation across both treatment and outcome models simultaneously, unlike traditional methods that consider only one model and are thus more likely to exclude important confounders with weak associations in the model under consideration. GLiDeR has desirable theoretical properties, and in simulation experiments outperforms doubly robust approaches which do not incorporate variable selection. It achieves similar efficiency with existing techniques which perform variable selection across both outcome and treatment models, but has substantial computational advantages over these approaches and allows for situations with $p > n$. Simulations suggest the largest gains in efficiency are achieved when the outcome is misspecified, a frequent occurrence in practice.

GLiDeR targets inference for the average causal treatment effect. Even though GLiDeR displays good performance in the simulation scenarios considered in Chapter 2, we caution that, like other model selection procedures, its finite sample performance at certain local alternatives can potentially be quite poor, reminiscent of Hodges' estimator (Leeb and Pötscher, 2008). While the validity of bootstrap intervals was not explored in Chapter 2, percentile bootstrap confidence intervals for the average causal effect had good coverage; how to adapt promising recent developments in post-selection inference to our setting is an

area of future research.

Alternatively, Chapter 3 proposes SSCE and BSSCE, two non-doubly robust Bayesian approaches for variable selection and estimation in causal inference that allow the variability of all parameters to be summarized from their posterior distributions. Both SSCE and BSSCE use spike and slab priors on the model coefficients to simultaneously model the outcome and treatment assignment. By simultaneously modeling the outcome and treatment assignment, like GLiDeR, the SSCE and BSSCE can identify confounding variables with weak associations in one model that may otherwise be ignored by a procedure using only that model, such as the Bayesian lasso on the outcome model. Furthermore, both SSCE and BSSCE show substantial improvements over BAC with $\omega = \infty$. SSCE aims to only adjust for covariates related to outcome or treatment assignment in order to minimize bias. Additionally, SSCE very infrequently includes irrelevant covariates that are unrelated to outcome and treatment assignment, which can greatly reduce variability and the MSE of the treatment effect estimator in finite samples compared to competing approaches such as BAC. On the other hand, BSSCE aims to reduce MSE of the treatment effect estimator by setting coefficients to zero for those covariates which if included in the outcome model would otherwise increase MSE of the treatment effect estimator. BSSCE provides an effective means of reducing MSE without having to data-adaptively choose a tuning parameter like ω in BAC. Furthermore, unlike BAC or other model averaging techniques, the proposed approaches performed well even when the number of covariates exceeded the sample size.

A linear model was assumed for the outcome throughout Chapters 2 and 3, but the proposed methods in those chapters could be extended to more complicated models with covariate-covariate interactions, polynomial transformations of covariates, covariate-treatment interactions, or smoothers of various kinds (e.g., penalized splines expressed as mixed linear models). However, additional work needs to be done for such extensions as the ideal grouping structure for interactions and transformations is unclear.

In Chapter 4, we move beyond average treatment effect estimation and propose TEHTrees, a new method for estimating heterogeneous causal treatment effects. Randomized studies are often used to make medical decisions and inform policy. Sometimes the results of randomized studies suggest that treatment effects are heterogeneous in a population, which may be used to pursue interventions or market treatments that target specific subgroups of the population. Recently, numerous methods have been proposed that aim to discover

treatment effect heterogeneity, including Causal Tree by Athey and Imbens (2016) (which was later extended to Causal Forests by Wager and Athey (2017)). While many of these approaches provide flexible subgroup identification, they typically do not address the issue of preserving Type I error rate. We propose TEHTrees, a new method for characterizing treatment effect heterogeneity that preserves the Type I error rate. TEHTrees uses decision trees to characterize treatment effect heterogeneity by utilizing matching within a conditional inference tree algorithm.

In simulation experiments, the Type I error rate using TEHTrees was below 0.05 (the pre-specified significance level) in all 23 considered scenarios, while the Type I error using Causal Tree was at least 0.15 in every scenario and even greater than 0.9 in many cases. Though Causal Tree has slightly greater power when there are continuous covariates, compared to TEHTrees, the power when using TEHTrees is actually greater than the power when using Causal Tree with binary covariates. When there is treatment effect heterogeneity, Causal Tree also tends to grow larger trees with more terminal nodes compared to TEHTrees, particularly with larger sample sizes. This makes it more difficult to infer the characteristics of groups that truly respond more (or less) to treatment when using Causal Tree compared to TEHTrees.

With binary covariates, treatment effect estimation was improved in our simulations when using TEHTrees compared to Causal Tree. However, Causal Tree displayed slightly better estimation properties than TEHTrees with continuous covariates, which was most likely due to greater variability in split points with TEHTrees. We conjecture that the variability in split points is larger with TEHTrees than with Causal Tree due to bias in the matching estimator. Decreasing bias in the matching estimator, or using an alternative approach to estimating the outcomes that are used as inputs in the conditional inference tree of TEHTrees, may therefore improve estimation of treatment effects with TEHTrees when there are continuous covariates.

TEHTrees is a flexible algorithm that allows for numerous modifications. A different matching algorithm can be implemented, an alternative prognostic score model can be fitted, other criterion can be used to find the splitting variable or its split point, and another estimation technique could be contrived and executed. The Bonferroni correction method that is used to find the splitting variable in TEHTrees (in Step 4a) is likely too conservative to detect small treatment effects when there are a large number of covariates

in the study. Alternative multiple comparison adjustment methods should be explored in the case when there are many covariates. Controlling the false discovery rate may also be a desirable alternative to multiple comparison adjustment methods when there are numerous covariates. TEHTrees also does not account for the variability in the matching estimator, so an extra step may be needed to control for the inflation of Type I error that could occur in situations when good matches are difficult to obtain.

Though many simulation scenarios were considered in Chapter 4, a larger variety of scenarios need to be considered. The prognostic score model, for example, was quite simple. Future research will increase the complexity of the prognostic score model and will also explore alternative effect sizes and treatment effect patterns (such as continuous interactions with treatment). Additionally, though we assume treatment is randomized (as in a clinical trial) throughout Chapter 4, TEHTrees could be extended for use with observational data; however, additional assumptions and steps in the algorithm would be required to achieve balance in the covariates, so this extension is left for future research.

Bibliography

Abadie, A. & Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects, *Econometrica*, **74**(1), 235–267.

Athey, S. & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences of the United States of America*, **113**(12), 7353–7360.

Aziz, F., Penupolu, S., Xu, X., and He, J. (2010). Lung transplant in end-staged chronic obstructive pulmonary disease (COPD) patients: a concise review, *Journal of Thoracic Disease*, **2**, 111-116.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables, *Journal of the American Statistical Association*, **57**, 33–55.

Blazère, M., Loubes, J. M., and Gamboa, F. (2014). Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension, *IEEE Transactions on Information Theory*, **60**, 2303–2318.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees, *Wadsworth Inc.*

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models, *American Journal of Epidemiology*, **163**, 1149–1156.

Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory, and Applications, *Berlin: Springer.*

Cefalu, M., Dominici, F., Arvold, N. D., and Parmigiani, G. (2017). Model averaged double robust estimation, *Biometrics*, **73**(2), 410–421.

Chatterjee, A & Lahiri, SN (2011), Bootstrapping lasso estimators, *Journal of the American Statistical Association*, **106**(494), 608–625.

Cochran, W. & Rubin, D. (1973). Controlling bias in observational studies: a review, *The Indian Journal of Statistics, Series A*, **35**(4), 417–446.

Crainiceanu, CM, Dominici, F, & Parmigiani, G (2008), Adjustment uncertainty in effect estimation, *Biometrika*, **95**(3), 635–651.

de Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect, *Biometrika*, **98**, 861–875.

Ertefaie, A., Asgharian, M., and Stephens, D. A. (2015). Variable selection in causal inference using a simultaneous penalization method, *arXiv preprint arXiv:1511.08501.*

Fan, G. Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A Decision Tree Approach, *Urban Studies*, **43**(12), 2301–2315.

Faries, D. E., Chen, Y., Lipkovich, I., Zagar, A., Liu, X., Obenchain, R. L. (2013). Local control for identifying subgroups of interest in observational research: persistence of treatment for major depressive disorder, *International Journal of Methods in Psychiatric Research*, **22**(3), 185–194.

Feller, A. and Holmes, C. (2009). Beyond toplines: heterogeneous treatment effects in randomized experiments, *Working Paper*, Columbia University.

Gass, K., Klein, M., Chang, H. H., Flanders, W. D., and Strickland, M. J. (2014). Classification and regression trees for epidemiologic research: an air pollution example, *Environmental Health*, **13**(1), 13–17.

Green, D. and Kern, H. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees, *Public Opinion Quarterly*, **76**(3), 491–511.

Hansen, B. (2008). The prognostic analogue of the propensity score, *Biometrika*, **95**(2), 481–488.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework, *Journal of Computational and Graphical Statistics*, **15**(3), 651-674.

Johnson, AEW, Pollard, TJ, Shen, L, Lehman, LH, Feng, M, Ghassemi, M, Moody, B, Szolovits, P, Celi, LA, & Mark, RG (2016), MIMIC-III, a freely accessible critical care database, *Scientific Data*, **3**, Article number 160035.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale, *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 13*, 1168-1176.

Lee, J, Kothari, R, Ladapo, JA, Scott, DJ, & Celi, LA (2012), Interrogating a clinical database to study treatment of hypotension in the critically ill, *BMJ Open*, **2**:e000916, https://doi.org/10.1136/bmjopen-2012-000916.

Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator, *Journal of Econometrics*, **142**, 201–211.

Lefebvre, G, Atherton, J, & Talbot, D (2014), The effect of the prior distribution in the Bayesian Adjustment for Confounding algorithm, *Computational Statistics & Data Analysis*, **70**, 227-240.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study, *Statistics in Medicine*, **23**, 2937–2960.

Mebane, W.R. and Sekhon, J. (2011). Genetic optimization using derivatives: the rgenoud package for r, *Journal of Statistical Software*, **42**(11), 1–26.

Morgan, J. and Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, **58**(302), 415–434.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and Team, R. Core (2015). nlme: Linear and nonlinear mixed eects models, 1–122.

Pouliakis, A., Margari, C., Margari, N., Chrelias, C., Zygouris, D., Meristoudis, C., Panayiotides, I., Karakitsos, P. (2014). Using classification and regression trees, liquidbased

cytology and nuclear morphometry for the discrimination of endometrial lesions, *Diagnostic Cytopathology*, **42**(7), 582–591. .

R Core Team. (2015). R: A language and environment for statistical computing, Vienna, Austria. URL: R Foundation for Statistical Computing.

Raftery, AE, Madigan, D, & Hoeting, JA (1997), Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**(437), 179–191.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**(1), 41–55.

Rubin, D. (1973a). Matching to remove bias in observational studies, *Biometrics*, **29**(1), 159–183.

Rubin, D. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics*, **29**(1), 185–203.

Sekhon, J. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for r, *Journal of Statistical Software*, **42**(7), 1–52.

Sela, R. and Simonoff, J. (2012). Re-em trees: a data mining approach for longitudinal and clustered data, *Machine Learning*, **86**, 169–207.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward, *Statistical Science*, **25**(1), 1–21.

Stuart, E., Lee, B. and Leacy, F. (2013). Prognostic score-based balance measures for propensity score methods in comparative effectiveness research, *Journal of Clinical Epidemiology*, **66**, 80.

Tibshirani, R (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011) MICE: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*, **45**, 1–67.

van der Laan, M. J. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation, *The International Journal of Biostatistics*, **6**, Article 17.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner,*Statistical Applications in Genetics and Molecular Biology*, **6**, Article 25.

VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection, *Biometrics*, **67**, 1406–1413.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference, *Statistical Methods in Medical Research*, **21**, 7–30.

Venkatasubramaniam, A., Wolfson, J., Mitchell, N., Barnes, T., Meghan JaKa, M., and French, S. (2017). Decision trees in epidemiological research, *Emerging Themes in Epidemiology*, **14**(11).

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty, *Biometrics*, **68**, 661–671.

Wager, S. & Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, *Journal of the American Statistical Association*, 1–15.

Xu, X & Ghosh, M (2015), Bayesian variable selection and estimation for group lasso, *Bayesian Analysis*, **10**(4), 909-936

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalized learning problems, *Statistics and Computing*, **25**, 1129–1141.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Methodological)*, **68**, 49–67.

Zigler, CM & Dominici, F (2014), Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects, *Journal of the American Statistical Association*, **109**(505), 95–107.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.